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 Самойличенко А.А.

Москва, 2023

**Содержание**

Введение3

1. Аналитическая часть4

1.1. Постановка задачи4

1.2. Описание используемых методов8

1.3. Разведочный анализ данных11

2. Практическая часть17

2.1. Предобработка данных17

2.2. Разработка и обучение модели21

2.3. Написать нейронную сеть, которая будет рекомендовать соотношение матрица. 25

2.4. Разработка приложения31

2.5. Создание удаленного репозитория и загрузка результатов работы на него.33

Заключение34

Список использованной литературы35

**Введение**

В последние годы сфера материаловедения стала одной из наиболее активно развивающихся отраслей, в которой большое внимание уделяется разработке новых материалов с улучшенными свойствами. Однако, создание новых материалов может быть дорогостоящим и времязатратным процессом, поэтому актуальным является разработка эффективных методов прогнозирования конечных свойств новых материалов на основе их состава и структуры.

В рамках курса "Data Science" в МГТУ имени Н.Э. Баумана было проведено исследование применения методов машинного обучения и создания нейронной сети для прогнозирования свойств композиционных материалов на основе данных о их составе и структуре. В данной выпускной квалификационной работе рассмотрены различные методы обработки и анализа данных, используемые в процессе построения моделей прогнозирования, а также проанализированы результаты применения различных алгоритмов машинного обучения и нейронных сетей для решения данной задачи.

Целью данной работы является разработка эффективной модели прогнозирования конечных свойств новых композиционных материалов на основе данных об их составе и структуре. Для достижения этой цели в работе рассматриваются различные подходы к анализу данных и методы машинного обучения, а также проводится сравнительный анализ результатов работы различных моделей.

В ходе работы были использованы данные о композиционных материалах, содержащие информацию об их составе и структуре, а также о конечных свойствах, которые необходимо было прогнозировать. Были проведены эксперименты с различными алгоритмами машинного обучения, в том числе с использованием нейронных сетей, деревьев решений, метода опорных векторов и случайных лесов. Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Результаты работы могут быть использованы в дальнейших исследованиях в области материаловедения, а также могут быть применены в производственных условиях для оптимизации процесса разработки новых материалов с требуемыми свойствами.

**1.** **Аналитическая часть**

**1.1. Постановка задачи.**

Созданные прогнозные модели на основе машинного обучения и нейронных сетей помогут сократить количество проводимых испытаний для композиционных материалов, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

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.

Датасет представлен с помощью 13 переменных, которые говорят о тех или иных свойствах данного композиционного материала:

* Соотношение матрица-наполнитель;
* Плотность, кг/м3;
* Модуль упругости, Гпа;
* Количество отвердителя, м.%;
* Содержание эпоксидных групп, %\_2;
* Температура вспышки, С\_2;
* Поверхностная плотность, г/м2;
* Модуль упругости при растяжении, ГПа;
* Прочность при растяжении, Мпа;
* Потребление смолы, г/м2;
* Угол нашивки, град;
* Шаг нашивки;
* Плотность нашивки.



Рисунок 1. Датасет.

Датасет со свойствами композитов представлен двумя файлами (X\_bp.xlsx и X\_nup.xlsx) после объединения которых (по индексу тип объединения INNER) в соответствии с поставленной задачей его размерность представляет собой 1023 строк и 13 столбцов. В данном датасете отсутствуют пропущенные значения и дубликаты, но присутствуют выбросы. Графики приведённые в данной работе показывают на наличие выбросов во всех признаках кроме 'Угол нашивки, град'. В основном распределения данных близки к нормальному, за исключением столбца "Угол нашивки, град" который представлен двумя значениями, которые были впоследствии заменены на значения 0 и 1.

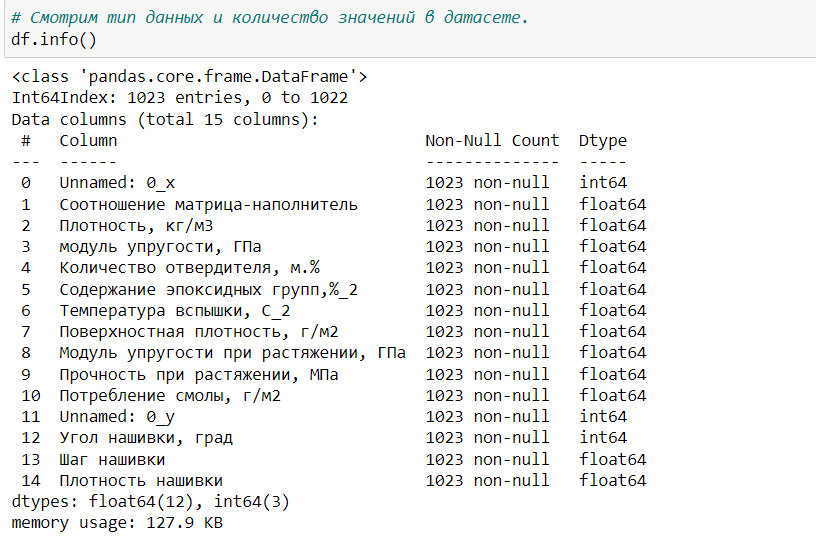


Рисунок 2. Количество и тип значений

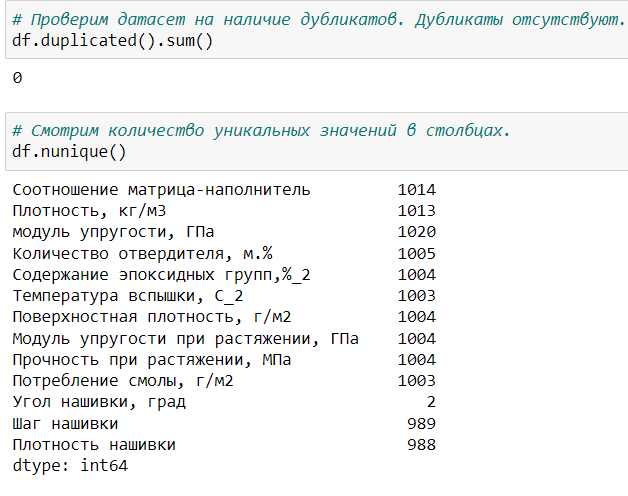


Рисунок 3. Дубликаты и количество уникальных значений

Для оценки корреляции между признаками была выбрана корреляция Пирсона. Корреляция Пирсона — это метод параметрической статистики, позволяющий определить наличие или отсутствие линейной связи между двумя количественными показателями, а также оценить ее тесноту и статистическую значимость. Также для наглядности зависимости между признаки была построена тепловая карта на которой заметно что корреляции между признаками не наблюдается.

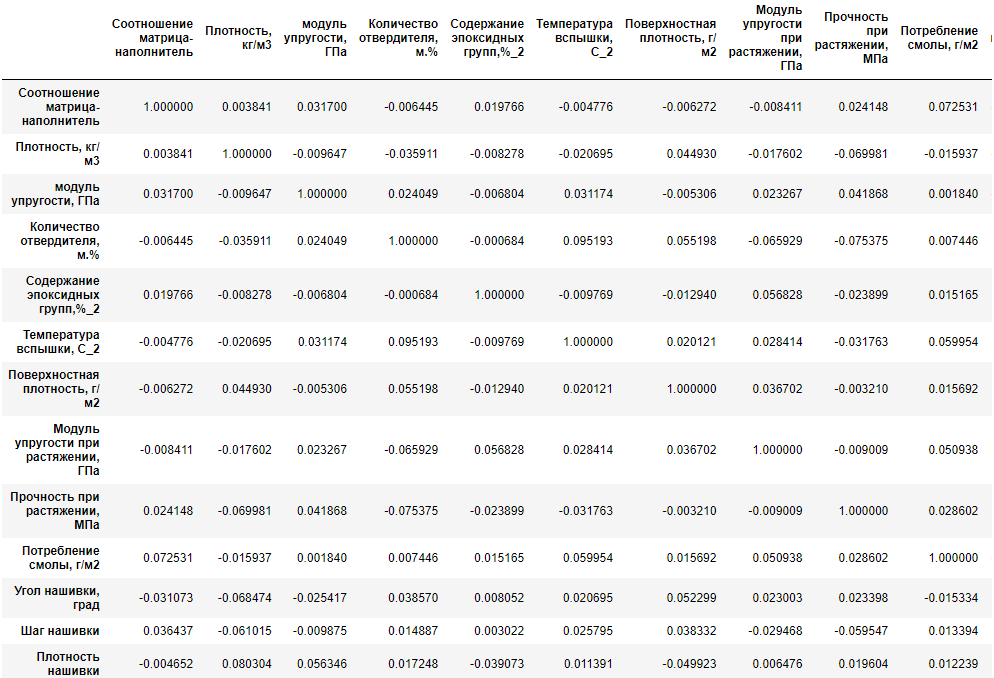


Рисунок 4. Корреляция Пирсона

В качества задачи для данной выпускной квалификационной работы ставится обучение алгоритма машинного обучения, который будет определять значения: “Модуль упругости при растяжении, ГПа" и “ Прочность при растяжении, МПа ”, а также создание нейронной сети, которая будет рекомендовать “Соотношение матрица-наполнитель”.

**1.2. Описание используемых методов**

Для решения данной задачи регрессии было использовано несколько методов машинного обучения:

* Lasso - Лассо (Лассо-регрессия);
* LinearRegression - Линейная регрессия;
* Ridge - Гребневая регрессия;
* DecisionTreeRegressor - Регрессионное дерево решений;
* GradientBoostingRegressor - Градиентный бустинг регрессии;
* RandomForestRegressor - Случайный лес регрессии;
* SVR - Метод опорных векторов для регрессии;
* BayesianRidge - Байесовская линейная регрессия;
* KernelRidge - Ядерная регрессия.

Лассо (Лассо-регрессия) — это метод регуляризации линейной регрессии, который использует L1-регуляризацию. Он работает путем добавления штрафа за сумму абсолютных значений коэффициентов регрессии, что приводит к сокращению некоторых коэффициентов до нуля, что позволяет выполнять отбор признаков.

Линейная регрессия — это простой метод машинного обучения для построения линейной модели, которая моделирует связь между входными признаками и выходными целевыми значениями. Он работает путем минимизации суммы квадратов ошибок между прогнозируемыми и реальными значениями.

Гребневая регрессия (Ridge) — это метод регуляризации линейной регрессии, который использует L2-регуляризацию. Он работает путем добавления штрафа за квадраты значений коэффициентов регрессии, что ограничивает их значения и уменьшает переобучение.

Дерево решений регрессии (Decision Tree Regression) — это метод нелинейной модели, который использует дерево решений для моделирования зависимости между признаками и целевой переменной. Дерево решений представляет собой структуру данных, которая состоит из узлов и листьев. Узлы представляют собой разделение признаков на две или более ветви, а листья — это конечные значения, которые являются предсказанными значениями для данного наблюдения. Дерево решений строится по обучающей выборке путем рекурсивного разделения на подгруппы, каждый раз выбирая признак, который максимально снижает дисперсию (или уменьшает ошибку).

Градиентный бустинг регрессии — это метод ансамблирования множества слабых моделей регрессии, чтобы получить более сильную модель. Он работает путем последовательного добавления слабых моделей, которые корректируют ошибки предыдущих моделей.

Случайный лес регрессии — это метод ансамблирования множества деревьев решений, чтобы получить более сильную модель. Он работает путем построения множества деревьев решений, которые каждое используют подмножество признаков и данных, и затем усредняет результаты прогнозирования всех деревьев.

Эластичная сеть регрессии (ElasticNetCV) - метод регрессии, который является комбинацией Лассо-регрессии и Гребневой регрессии. Он минимизирует сумму квадратов ошибок и включает в функционал потерь штрафы как на L1-, так и на L2-нормы коэффициентов регрессии. Это позволяет улучшить результаты в случаях, когда в данных присутствуют коррелированные признаки.

Метод опорных векторов для регрессии (SVR) – это метод регрессии, который использует максимальный зазор между опорными векторами для построения линейной или нелинейной модели регрессии. Он минимизирует сумму квадратов ошибок и добавляет ограничения на значения коэффициентов регрессии в зависимости от расстояний до опорных векторов.

Байесовская линейная регрессия (Bayesian Ridge) - это метод регрессии, который использует байесовский подход для оценки коэффициентов регрессии и шума в данных. Он минимизирует сумму квадратов ошибок и моделирует коэффициенты регрессии как случайные переменные с нормальным распределением.

Ядерная регрессия (Kernel Ridge) — это метод регрессии, который использует ядерные функции для построения нелинейной модели регрессии. Он минимизирует сумму квадратов ошибок и добавляет штраф на сложность модели за счет добавления к функционалу потерь L2-нормы коэффициентов регрессии.

Таблица 1 - Сравнительная таблица с указанием априорных предпосылок к работоспособности каждого метода

|  |  |
| --- | --- |
| **Метод машинного обучения** | **Априорные предпосылки** |
| Лассо (Лассо-регрессия) | Наличие признаков, которые не являются важными для модели; мультиколлинеарность между признаками. |
| Линейная регрессия | Линейная зависимость между входными признаками и выходными целевыми значениями |
| Гребневая регрессия (Ridge) | Мультиколлинеарность между признаками; необходимость сокращения размерности признаков |
| Регрессионное дерево решений | Нелинейная зависимость между входными признаками и выходными целевыми значениями |
| Градиентный бустинг регрессии | Наличие множества слабых моделей, которые могут быть объединены в более сильную модель |
| Случайный лес регрессии | Наличие множества слабых моделей, которые могут быть объединены в более сильную модель |
| Эластичная сеть регрессии (ElasticNetCV) | Эффективна в случаях, когда в данных присутствуют коррелированные признаки. |
| Метод опорных векторов для регрессии (SVR) | Нелинейная зависимость между входными признаками и выходными целевыми значениями; необходимость нахождения гиперплоскости, которая разделяет данные на две части с наибольшим зазором |
| Байесовская линейная регрессия  (Bayesian Ridge) | Эффективна при работе с данными, когда нужно моделировать неопределенность в коэффициентах регрессии и шуме в данных. |
| Ядерная регрессия  (Kernel Ridge) | Нелинейная зависимость между входными признаками и выходными целевыми значениями; необходимость применения ядерной функции для построения нелинейных границ |

**1.3. Разведочный анализ данных**

Разведочный анализ данных — это анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

В данном проекте были использованы следующие методы разведочного анализа данных:

* Описательная статистика данного датасета.
* Визуальный анализ гистограмм
* Визуальный анализ диаграмм размаха («ящик с усами»)
* Проверка нормальности распределения по критерию Пирсона
* Анализ попарных графиков рассеяния переменных
* Корреляционный анализ c целью поиска коэффициентов

Для просмотра статистических данных для в каждого столбца используется метод describe( ). Этот метод показывает количество строк в столбце - count, среднее значение столбца - mean, столбец стандартное отклонение - std, минимальные (min) и максимальные (max) значения, а также границу каждого квартиля - 25%, 50% и 75%. Любые значения NaN автоматически пропускаются.

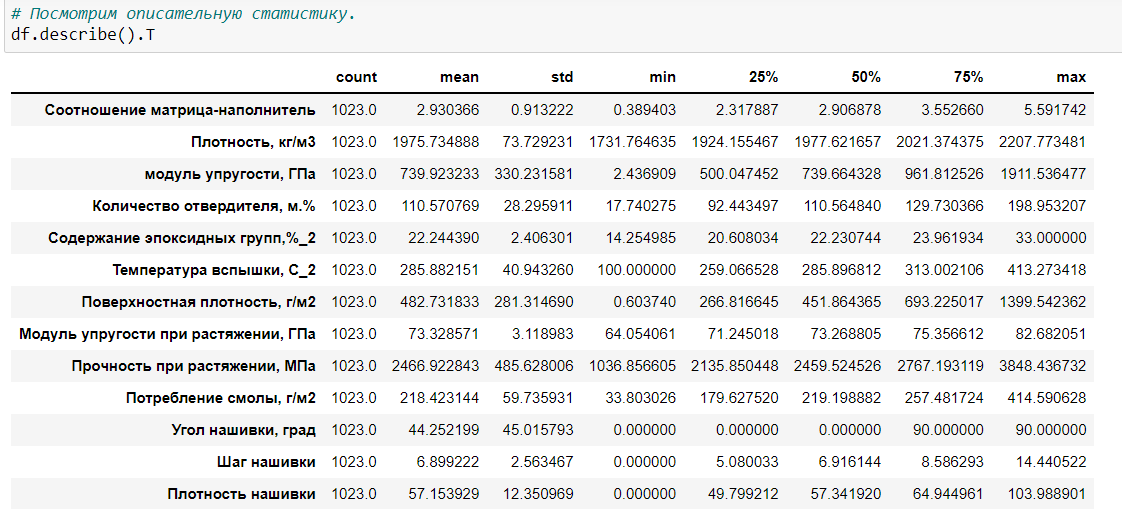


Рисунок 5. Описательная статистика

Для просмотра количества уникальных значений в датасете используем метод nunique().

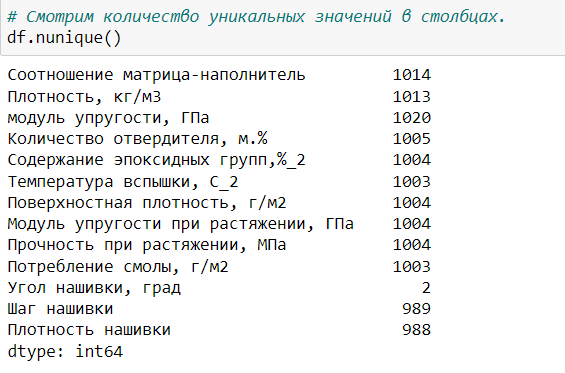


Рисунок 6. Количество уникальных значений в столбцах.

Для просмотра количества дубликатов в датасете используем метод duplicated().

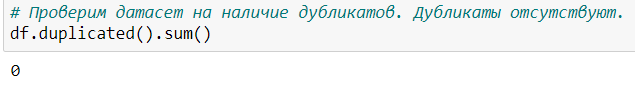


Рисунок 7. Проверка на наличие дубликатов

Для оценки величины и характера разброса данных построены гистограммы распределения для каждого из признаков.

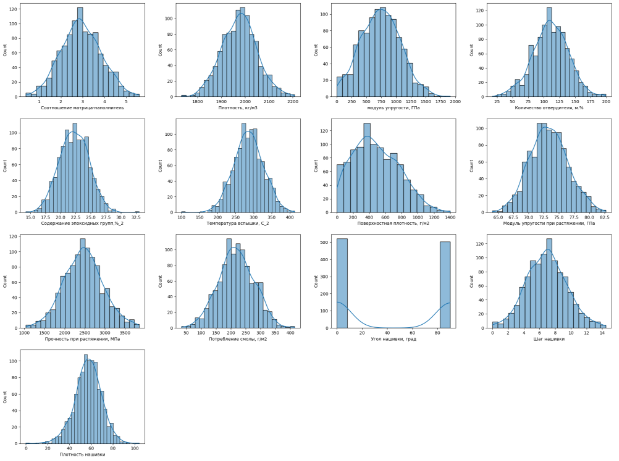


Рисунок 8. Гистограммы распределения

Гистограммы показывают что в основном распределения данных близки к нормальному, за исключением столбца с данными “Угол нашивки, град”, который представлен двумя значениями.

Для проверки датасета на наличие выбросов был проведён визуальный анализ диаграмм размаха. Диаграммы размаха показывают медиану, верхний квартиль, нижний квартиль, межквартильный размах, выбросы. На «ящиках с усами» хорошо заметны наличия выбросов в данном датасете.

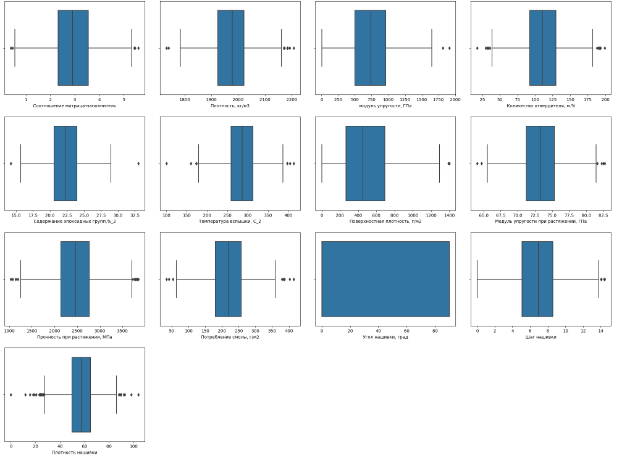


Рисунок 9. Диаграмма размаха

Так же был проведён анализ попарных графиков рассеяния переменных.

Попарный график помогает нам визуализировать распределение отдельных переменных, а также взаимосвязи между двумя переменными. Это отличный метод определения тенденций между переменными для последующего анализа. Если точки на графике расположены вдоль прямой линии, это может указывать на линейную связь между переменными. Если точки расположены в случайном порядке на графике, это может указывать на отсутствие связи между переменными или на то, что связь между переменными сложна и не может быть описана простой линейной связью. Также попарный график рассеяния может помочь идентифицировать выбросы, что являются отдельными наблюдениями, которые выделяются на графике и могут оказывать большое влияние на результаты анализа.

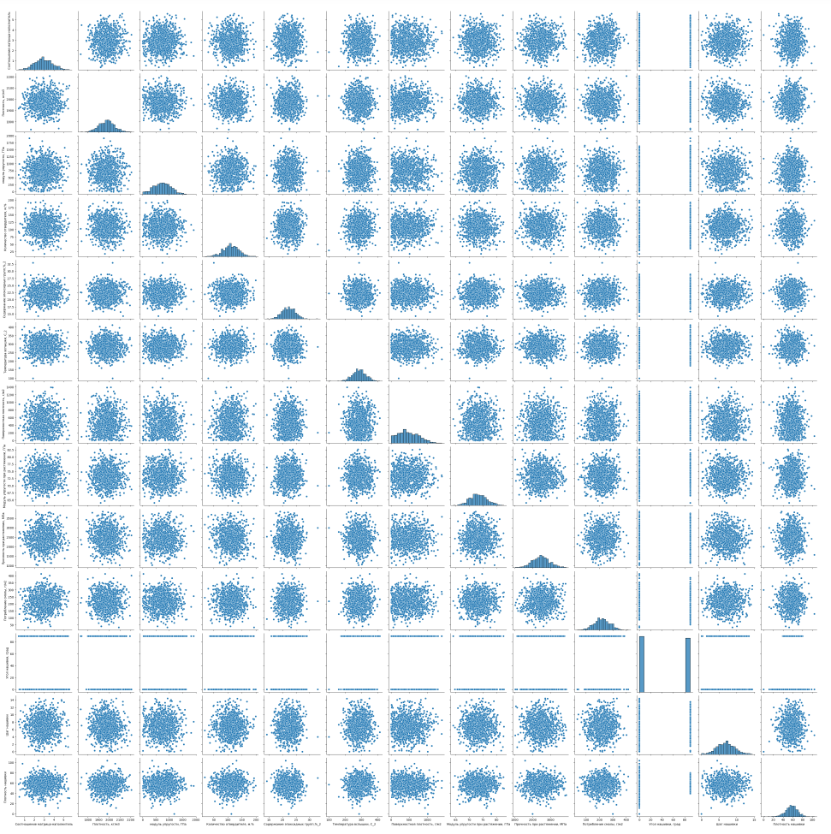


Рисунок 10. Попарный график рассеяния

График оценки плотности показывает, как вероятность распределена по значению непрерывной случайной величины. Он может использоваться для визуализации формы распределения, а также для получения информации о вероятности различных значений этой случайной величины.

График оценки плотности может использоваться для выявления выбросов, определения типа распределения, сравнения распределений между собой, оценки вероятности событий и принятия решений на основе статистических данных.

Оценка плотности ядра показывает, что значения переменных находятся в разных диапазонах поэтому требуется нормализация данных.

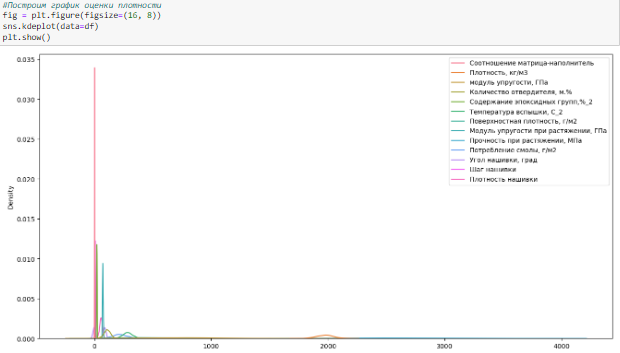


Рисунок 11. График оценки плотности распределения

Для визуализации коэффициентов корреляции и определения наличия между переменными зависимости, была построена тепловая карта коэффициентов корреляции методом Пирсона.

Корреляция Пирсона — это статистический показатель, который используется для измерения степени линейной связи между двумя переменными. Она измеряет силу и направление линейной связи между двумя непрерывными переменными.

Коэффициент корреляции Пирсона принимает значения от -1 до 1. Значение -1 указывает на полную обратную линейную связь, значение 1 указывает на полную прямую линейную связь, а значение 0 указывает на отсутствие линейной связи между переменными.

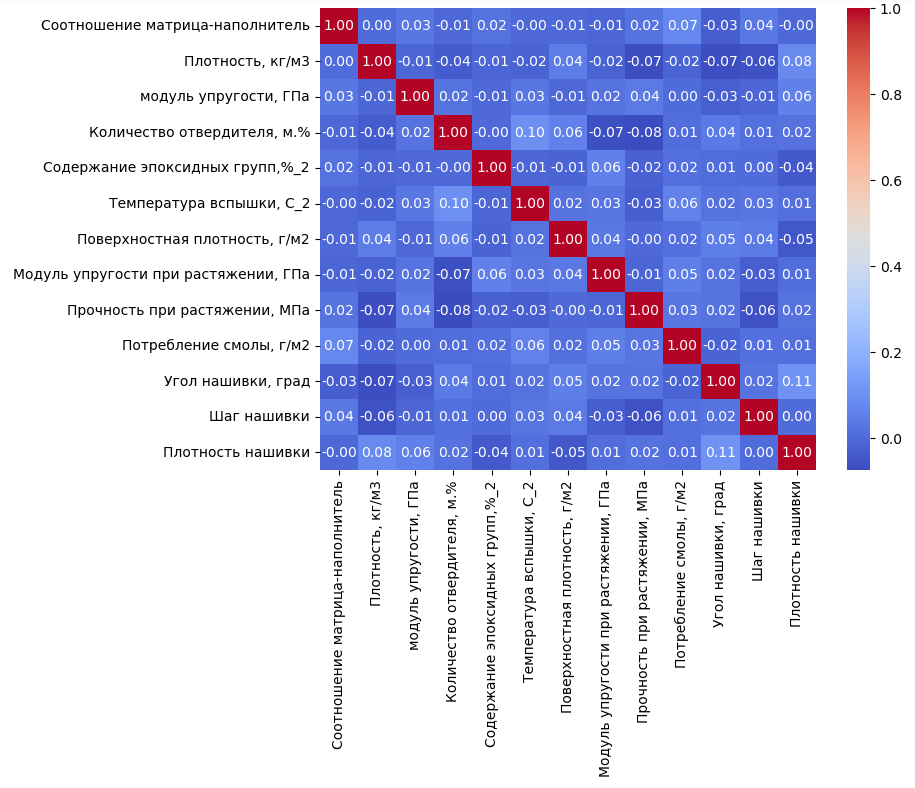


Рисунок 12. Тепловая карта

Корреляции между переменными практически не наблюдается.

**2. Практическая часть**

**2.1 Предобработка данных**

Для удаления выбросов в датасете был выбран метод исключения выбросов с помощью правила трёх сигм.

Правило трех сигм — это статистический метод для определения выбросов в наборе данных. Этот метод основан на стандартном отклонении данных от среднего значения. Согласно правилу трех сигм, большинство данных должны находиться в пределах трех стандартных отклонений от среднего значения. Используя эту концепцию, можно определить, какие данные являются выбросами.

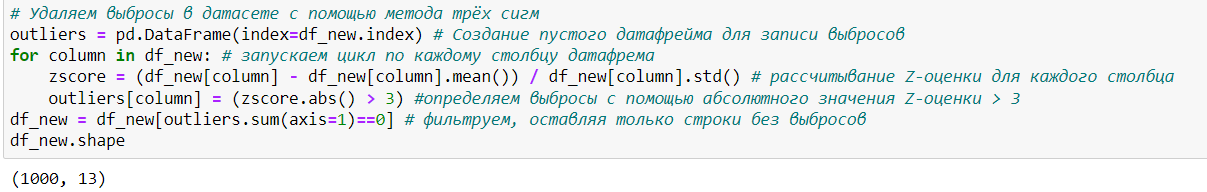


Рисунок 13. Удаления выбросов

Очищенный от выбросов датасет содержит 1000 строк. С помощью диаграмм размаха проверяем на сколько хорошо были удалены выбросы в датасете.

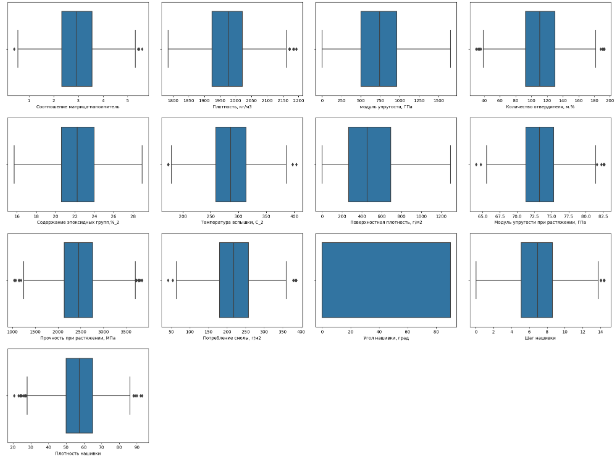


Рисунок 14. Диаграмма рассеивания

Так как признак «Угол нашивки» представлен только двумя значениями: “0” и “90”, то эти значения были заменены на 0 и 1.

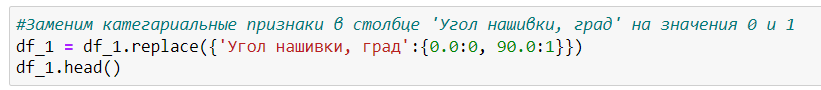


Рисунок 15. Замена категориальных признаков

Для приведения всех признаков в датасете к одинаковой шкале была сделана нормализация с помощью метода MinMaxScaler из библиотеки sklearn.

Нормализация — это процесс приведения данных к определенному масштабу или диапазону значений. Нормализация может использоваться для приведения различных признаков в датасете к единому масштабу, чтобы они имели одинаковый вклад в результаты анализа или моделирования.

Нормализация данных является важным шагом в подготовке данных для машинного обучения и может улучшить результаты алгоритмов, повысить их скорость и устойчивость к шуму.

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

Формула для MinMax-нормализации следующая:

где:

x\_scaled - отмасштабированное значение признака;

x - оригинальное значение признака;

x\_min - минимальное значение признака в наборе данных;

x\_max - максимальное значение признака в наборе данных.

Процесс нормализации MinMaxScaler включает в себя следующие шаги:

* Определение минимального и максимального значений признака в наборе данных.
* Применение формулы MinMax-нормализации для каждого значения признака.
* Получение отмасштабированных значений признаков в диапазоне от 0 до 1.

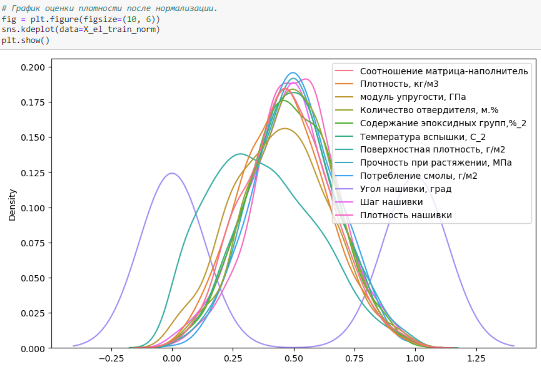


Рисунок 16. График оценки плотности распределения после нормализации

Описательная статистика тренировочной выборки до нормализации.

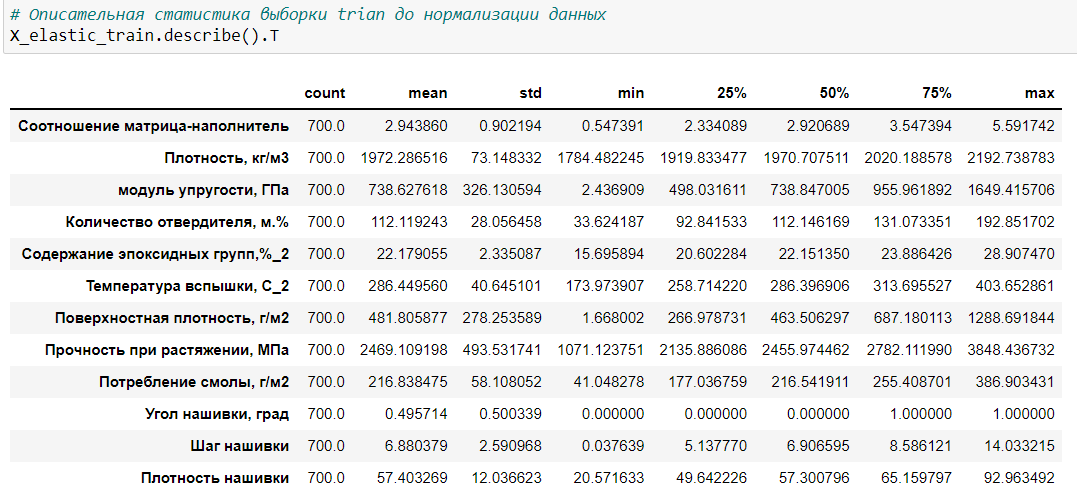


Рисунок 17. Описательная статистика до нормализации

Описательная статистика тренировочной выборки после нормализации.

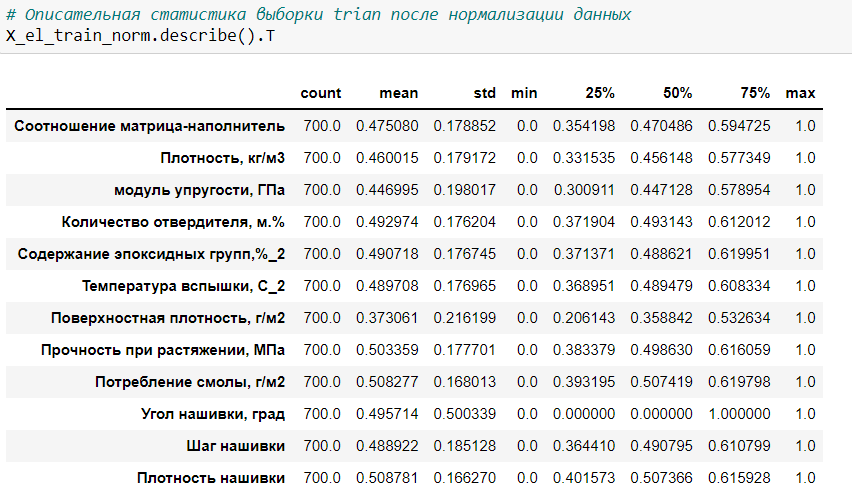


Рисунок 17. Описательная статистика после нормализации

* 1. **Разработка обучение и тестирование модели**

Для решения поставленной задачи данные были разделены на обучающую и тестовую выборки в соотношении 70/30. Для каждой модели был создан словарь с гиперпараметрами и с помощью поиска по сетке с перекрестной проверкой были найдены лучшие гиперпараметры для каждой модели. Оценка качества работы каждой из моделей выполнялась с помощью вычисления коэффициента детерминации (R²), среднеквадратичной ошибки (RMSE) и средней абсолютной ошибка (MAE).

Для прогнозирования модуля упругости при растяжении и прочности при растяжении были использованы следующие методы машинного обучения:

* Лассо-регрессия (Lasso);
* Линейная регрессия (LinearRegression);
* Гребневая регрессия (Ridge);
* Регрессионное дерево решений (DecisionTreeRegressor);
* Градиентный бустинг регрессии (GradientBoostingRegressor);
* Случайный лес регрессии (RandomForestRegressor);
* Эластичная сеть регрессии (ElasticNetCV);
* Метод опорных векторов для регрессии (SVR);
* Байесовская линейная регрессия (BayesianRidge);
* Ядерная регрессия (KernelRidge).

Коэффициент детерминации, имеющий отрицательные значения близкие к нулю, говорит о том, что результат использования моделей не точнее использования для прогноза среднего значения прогнозируемого параметра.

Наилучший результат для “Модуль упругости при растяжении” с подобранными гиперпараметрами на тренировочной выборке показала модель KernelRidge.

Таблица 2 - Показатели моделей на тренировочной выборке, предсказывающих модуль упругости при растяжении.

| **Метод машинного обучения** | **R2** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Lasso | -0.014721 | -3.118528 | -2.502596 |
| LinearRegression | -0.020283 | -3.126736 | -2.502399 |
| Ridge | -0.017374 | -3.122493 | -2.500838 |
| DecisionTreeRegressor | -0.025911 | -3.136856 | -2.512126 |
| GradientBoostingRegressor | -0.050991 | -3.173937 | -2.533804 |
| RandomForestRegressor | -0.030986 | -3.143604 | -2.517188 |
| ElasticNetCV | -0.013461 | -3.117018 | -2.498687 |
| SVR | -0.016112 | -3.119945 | -2.485916 |
| BayesianRidge | -0.013430 | -3.116968 | -2.499407 |
| KernelRidge | -0.012441 | -3.115224 | -2.492515 |

Все модели для расчёта “ модуля упругости при растяжении” на тестовой выборке показали неудовлетворительные результаты, коэффициент детерминации у всех моделей показал отрицательное значение. Лучшие показатели на тестовой выборке показывает модель BayesianRidge и ElasticNetCV.

Таблица 3 - Показатели моделей на тестовой выборке, предсказывающих модуль упругости при растяжении.

| **Метод машинного обучения** | **R2** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Lasso | -0.021946 | 3.104460 | 2.512111 |
| LinearRegression | -0.042239 | 3.135130 | 2.542403 |
| Ridge | -0.036315 | 3.126208 | 2.533229 |
| DecisionTreeRegressor | -0.065842 | 3.170432 | 2.551608 |
| GradientBoostingRegressor | -0.121470 | 3.252115 | 2.595599 |
| RandomForestRegressor | -0.058812 | 3.159959 | 2.527859 |
| ElasticNetCV | -0.021981 | 3.104512 | 2.510919 |
| SVR | -0.062883 | 3.166028 | 2.556054 |
| BayesianRidge | -0.021522 | 3.103815 | 2.510063 |
| KernelRidge | -0.031442 | 3.118849 | 2.524650 |

Единственная модель для расчёта “Прочности при растяжении” на тренировочной выборке, у которой коэффициент детерминации стал положительным - стала модель SVR (Метод опорных векторов).

Таблица 4 - Показатели моделей на тренировочной выборке, предсказывающих “Прочности при растяжении”.

| **Метод машинного обучения** | **R2** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Lasso | -0.011016 | -490.494300 | -389.793813 |
| LinearRegression | -0.016506 | -491.738397 | -391.673424 |
| Ridge | -0.014197 | -491.217848 | -390.996144 |
| DecisionTreeRegressor | -0.052261 | -500.563981 | -397.438065 |
| GradientBoostingRegressor | -0.008193 | -489.907915 | -391.846502 |
| RandomForestRegressor | -0.007580 | -489.809585 | -389.693131 |
| ElasticNetCV | -0.022941 | -493.539259 | -391.022774 |
| SVR | 0.003436 | -486.965371 | -385.813515 |
| BayesianRidge | -0.022944 | -493.539871 | -391.010971 |
| KernelRidge | -0.011806 | -490.780486 | -389.554120 |

Все модели для расчёта “Прочности при растяжении” на тестовой выборке показали неудовлетворительные результаты, коэффициент детерминации у всех моделей показал отрицательное значение.

Лучшие показатели на предсказанных значениях показали BayesianRidge и ElasticNetCV.

Таблица 5 - Показатели моделей на тестовой выборке, предсказывающих “Прочности при растяжении”.

| **Метод машинного обучения** | **R2** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Lasso | -0.019856 | 469.004096 | 365.853564 |
| LinearRegression | -0.041048 | 473.851843 | 370.003339 |
| Ridge | -0.034655 | 472.394687 | 368.941183 |
| DecisionTreeRegressor | -0.025416 | 470.280839 | 369.984701 |
| GradientBoostingRegressor | -0.030947 | 471.547480 | 371.453000 |
| RandomForestRegressor | -0.037358 | 473.011207 | 372.443349 |
| ElasticNetCV | -0.000928 | 464.631542 | 363.886617 |
| SVR | -0.049553 | 475.783433 | 368.535000 |
| BayesianRidge | -0.000928 | 464.631544 | 363.886619 |
| KernelRidge | -0.010346 | 466.812301 | 364.936087 |

**2.3 Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.**

Для решения поставленной задачи по рекомендации значения параметра “Соотношение матрица-наполнитель” были написаны нейросети с помощью MLPRegressor из библиотеки sklearn и с помощью Sequential из библиотеки Keras.

Для работы с нейросетью датасет разделен на новые тестовую и обучающую выборку в соотношении 70/30, с целевым параметром «Соотношение матрица/наполнитель».

Для подбора лучших параметров для MLPRegressor был использован метод GridSearchCV из библиотеки sklearn.

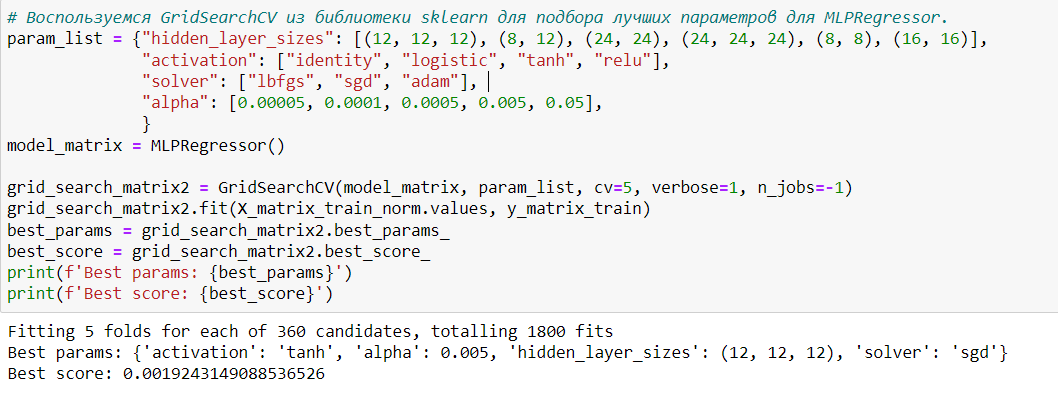


Рисунок 18. Подбор гиперпараметров

Преимущества многослойного перцептрона:

* Возможность изучать нелинейные модели.
* Возможность изучения моделей в режиме реального времени.
* К недостаткам многослойного персептрона (MLP) можно отнести:
* MLP со скрытыми слоями имеют невыпуклую функцию потерь, когда существует более одного локального минимума. Поэтому разные инициализации случайных весов могут привести к разной точности проверки.
* MLP требует настройки ряда гиперпараметров, таких как количество скрытых нейронов, слоев и итераций.
* MLP чувствителен к масштабированию функций.

Архитектура и параметры нейронной сети многослойного персептрона MLPRegressor:

* Последовательная модель нейронной сети.
* Модель состоит из трёх скрытых слоев, в каждом слое по 12 нейронов и выходного слоя с одним нейроном.
* Функция активации слоев выбран гиперболический тангенс (tanh). В качестве оптимизатора нейронной сети используется SGD.
* (стохастический градиентный спуск) с импульсом ускорения = 0.5.
* В нейронной сети используется функция ранней остановки обучения.
* Параметр регуляризации L2 (регуляризация весов), который уменьшает переобучение модели - 0.005.
* Максимальное количество итераций для обучения модели – 500.
* Для валидации будет использовано 30% обучающих данных.

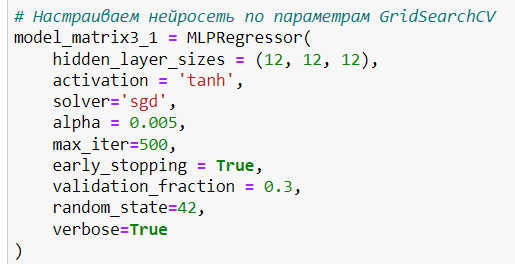


Рисунок 19. Настройка нейронной сети

В процессе обучения нейронной сети из библиотеки Keras сработала ранняя остановка обучения на 17-ой эпохе обучения, т.к. в течение 10-ти эпох не наблюдалось улучшения результата потерь на валидационной выборке.

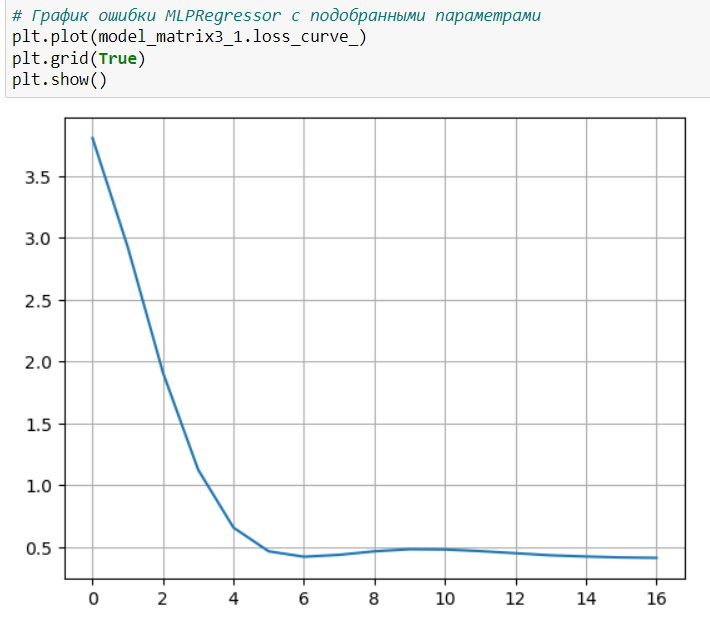


Рисунок 20. График ошибки.

Коэффициент детерминации для MLPRegressor составил: -0.019444.

Среднеквадратичная ошибка для MLPRegressor: 0.932634.

График прогнозных данных, полученных с помощью нейронной сети MLPRegressor.

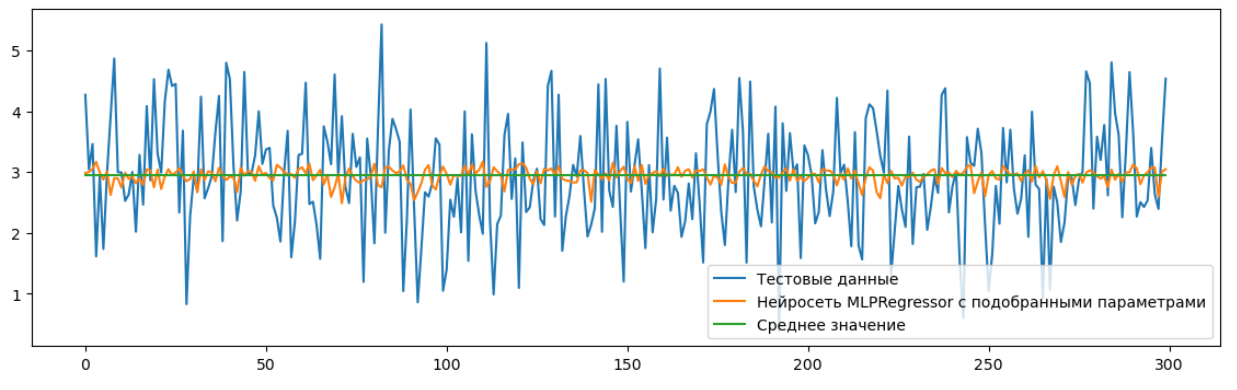


Рисунок 21. График прогнозных данных

Модель нейронной сети, созданной с помощью MLPRegressor показала неудовлетворительный результат. Коэффициент детерминации, имеющий значение близкое к нулю, говорит о том, что результат использования нейронной сети не точнее использования для прогноза среднего значения прогнозируемого параметра.

Архитектура и параметры нейронной сети из библиотеки Keras:

* Последовательная модель (Sequential) нейронной сети.
* Модель состоит из трёх скрытых слоев (Dense), с количеством нейронов, в которых равно 8, 16, 24 и выходного слоя с одним нейроном, так же в нейронной сети есть три слоя регуляризации (Dropout) со значениями (0.2, 0.3, 0.5).
* Функция активации слоев выбран гиперболический тангенс (tanh).
* В качестве оптимизатора нейронной сети используется SGD (стохастический градиентный спуск) со скоростью обучения 0,005, функцией потерь "среднеквадратическая ошибка" и метрикой оценки качества "среднеквадратичная ошибка".
* В нейронной сети используется функция ранней остановки обучения, если в течение пяти эпох не наблюдается улучшения потерь на валидационной выборке.
* Количество эпох обучения равно 100
* Для валидации будет использовано 30% обучающих данных.

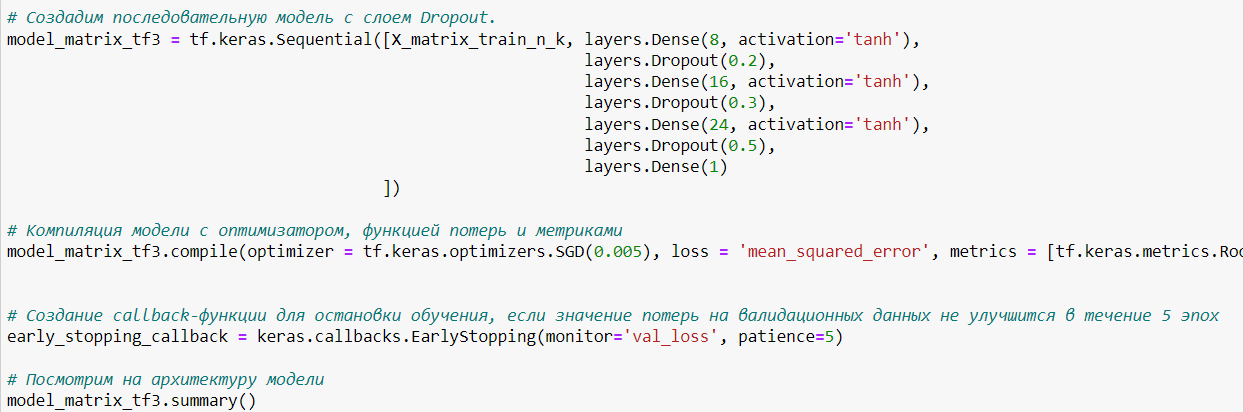


Рисунок 22. Нейронная сеть

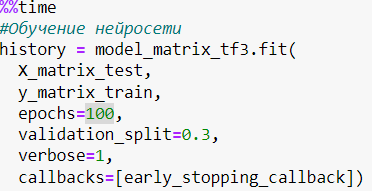


Рисунок 23. Обучение нейронной сети

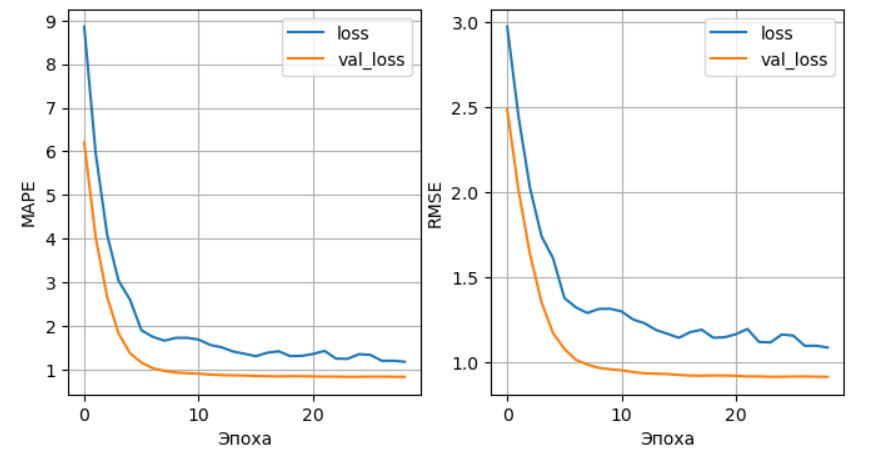


Рисунок 24. График ошибки

Коэффициент детерминации для последовательной нейронной сети из библиотеки Keras: -0.021756

Среднеквадратичная ошибка для последовательной нейронной сети из библиотеки Keras: 0.933692

График прогнозных значений, полученных с помощью последовательной нейронной сети из библиотеки Keras.

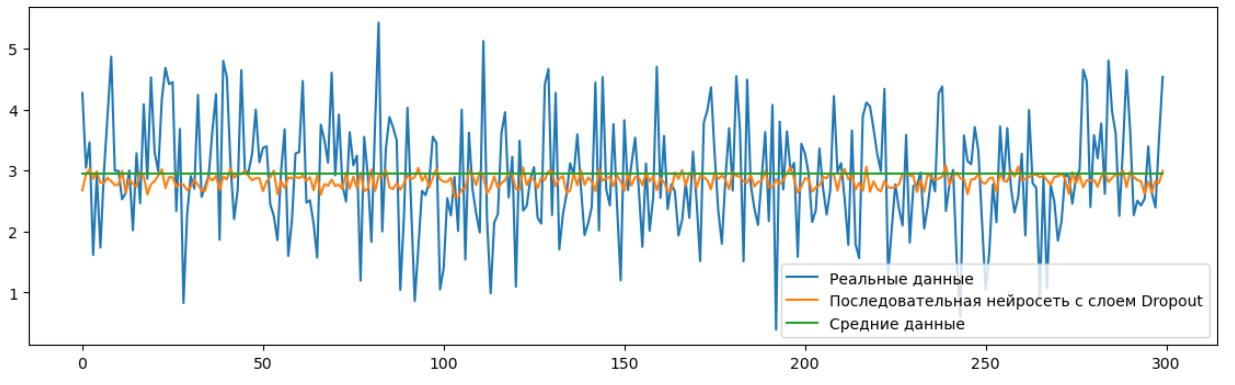


Рисунок 25. График прогнозных значений

Модель последовательной нейронной сети, созданной с помощью библиотеки Keras показала неудовлетворительный результат. Коэффициент детерминации, имеющий значение близкое к нулю, говорит о том, что результат использования нейронной сети не точнее использования для прогноза среднего значения прогнозируемого параметра.

Таблица 6 - Результаты работы нейронных сетей для предсказания “Соотношение матрица-наполнитель”

| **Нейронная сеть** | **R2** | **RMSE** |
| --- | --- | --- |
| MLPRegressor без подобранных гиперпараметров | -0.037495 | 0.940855 |
| MLPRegressor с подобранными гиперпараметрами | -0.019444 | 0.932634 |
| Последовательная нейросеть (Keras) | -0.059840 | 0.950933 |
| Последовательная нейросеть с callback (Keras) | -0.109839 | 0.973105 |
| Последовательная нейросеть с Dropout (Keras) | -0.021756 | 0.933692 |

**2.4 Web-приложение на Flask**

Веб-приложение для прогноза соотношения «Матрица-наполнитель». написано на языке программирования Python использованием библиотеки Flask, для написания шаблонов страниц был использован язык разметки HTML.

1. Для запуска приложения необходимо:
2. Запустить среду разработки Visual Studio Code.
3. Запустить приложение командой "python app.py".
4. Перейти по сгенерированной ссылке.
5. Нажать на кнопку "Спрогнозировать значение матрица-наполнитель"
6. Ввести значение для каждого параметра.
7. Нажать кнопку "Рассчитать"

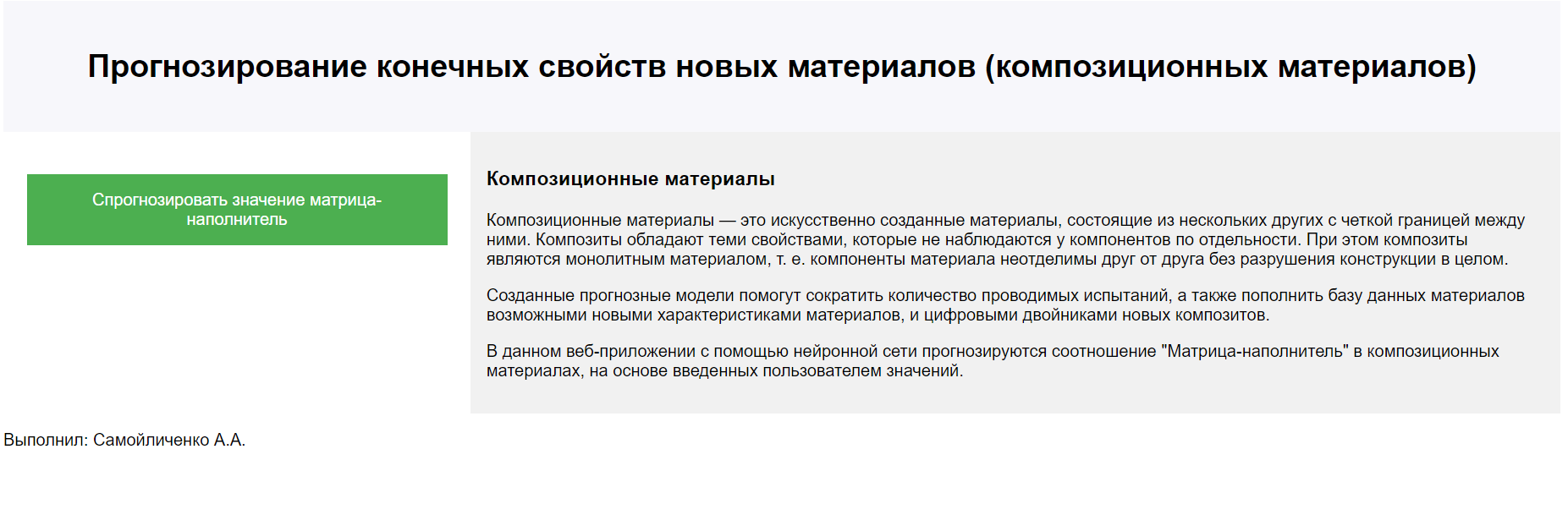


Рисунок 26. Первая страница приложения

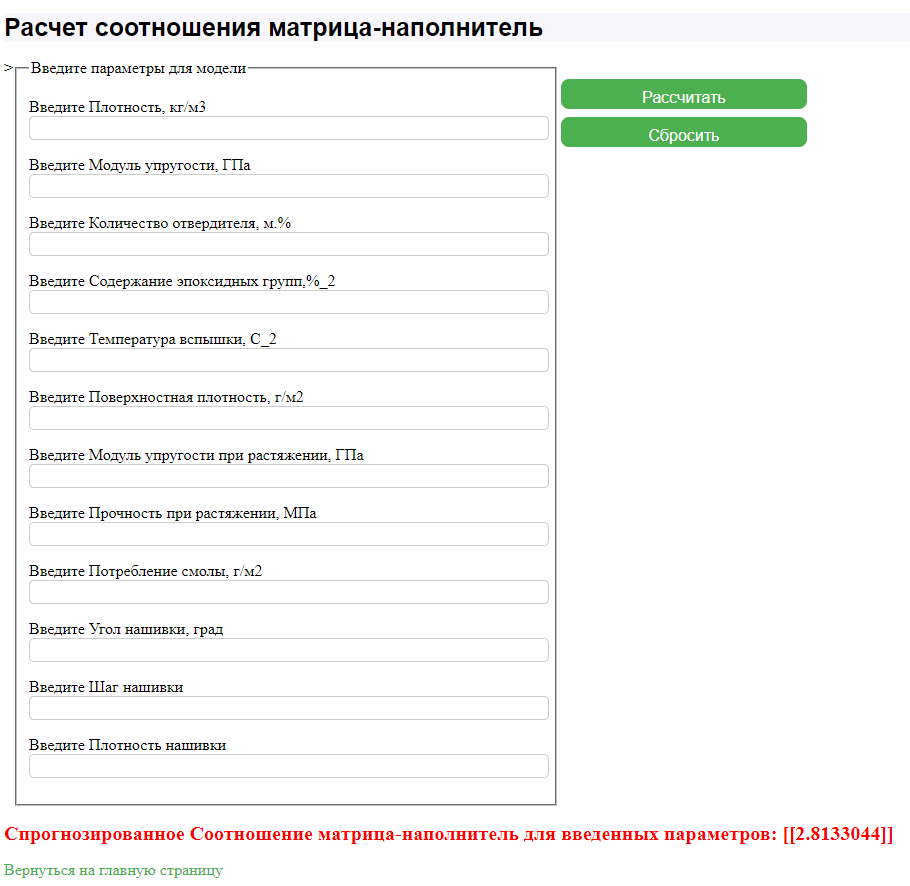


Рисунок 27. Вторая страница приложения

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

Ссылка на репозиторий в GitHub:

<https://github.com/andreigh89/vkr>

**Заключение**

В данной работе не удалось разработать эффективную модель прогнозирования конечных свойств новых композиционных материалов на основе данных об их составе и структуре, точность предсказанных значений практически не превосходило среднее значения. Возможно для улучшения работы предсказательных систем следует уделить больше внимания качеству и количеству признаков в датасете.

Тем не менее получен существенный практический опыт по анализу, визуализации и предобработке данных, созданию нейронных сетей и моделей машинного обучения, а также опыт по созданию веб-приложения на основе этих прогнозных моделей.

**Список использованной литературы**

1. Композиционные материалы: Справочник /Под. ред. В.В. Васильева, Ю.М.Тарнопольского. –М.: Машиностроение, 1990. –512 с.
2. Библиотека Keras - инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2018. - 294 с.
3. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
4. Платформа scikit-learn [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/ (дата обращения: 21.03.2023).
5. Библиотека Seaborn- Режим доступа: https://seaborn.pydata.org/. (дата обращения 22.03.2023)
6. Язык программирования Python- Режим доступа: https://www.python.org/. (дата обращения 12.03.2023)
7. Библиотека Pandas – Режим доступа: https://pandas.pydata.org/ (дата обращения 17.03.2023)
8. Библиотека Sklearn – Режим доступа: https://scikit-learn.org/stable/ (дата обращения 23.03.2023)
9. Библиотека Pandas- Режим доступа: https://pandas.pydata.org/. (дата обращения 22.03.2023)
10. Библиотека Matplotlib- Режим доступа: https://matplotlib.org/. (дата обращения 18.03.2023)
11. Библиотека Tensorflow: Режим доступа: https://www.tensorflow.org/. (дата обращения 24.03.2023)