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 Самсоненко Елена Евгеньевна

Москва, 2023

**Содержание**

[**Введение** 3](#_Toc130908375)

[**1.** **Аналитическая часть** 5](#_Toc130908376)

[**1.1.** **Постановка задачи** 5](#_Toc130908377)

[**1.2.** **Описание используемых методов** 5](#_Toc130908378)

[**1.3.** **Разведочный анализ данных** 14](#_Toc130908379)

[**2.** **Практическая часть** 20](#_Toc130908380)

[**2.1.** **Предобработка данных** 20](#_Toc130908381)

[**2.2.** **Разработка и обучение модели** 23](#_Toc130908382)

[**2.3.** **Тестирование модели** 26](#_Toc130908383)

[**2.4.** **Написать нейронную сеть, которая будет рекомендовать соотношение матрица.** 27](#_Toc130908384)

[**2.5.** **Разработка приложения** 30](#_Toc130908385)

[**2.6.** **Создание удаленного репозитория** 32](#_Toc130908386)

[**Заключение** 33](#_Toc130908387)

[**Библиографический список** 34](#_Toc130908388)

# **Введение**

Композитные материалы - это искусственно созданные новые материалы, которые состоят из двух или более материалов с уникальными свойствами, которые отличаются от их исходных компонентов. Композиты широко используются в различных промышленных отраслях, таких как авиация, космическая техника, автомобилестроение, строительство, судостроение, радиоэлектроника, производство бытовых предметов и спортивного инвентаря.

Существует два основных способа определения характеристик композитных материалов: физические испытания на образцах материала и прогнозирование свойств на основе моделирования. В данной работе мы сосредоточимся на втором подходе. Он основан на симуляции характеристик представительного элемента объема композита на основе данных о начальных свойствах компонентов, таких как количество связующего, наполнителя, температурный режим отверждения и другие параметры. Цель аналитической записки заключается в прогнозировании конечных свойств получаемых композиционных материалов на основе этих входных данных.

Одним из преимуществ использования регрессионных моделей и нейронных сетей в определении характеристик композитов является возможность быстрого и точного прогнозирования свойств новых материалов на основе небольшого количества данных. Это позволяет создавать новые и улучшать существующие композитные материалы с оптимальными свойствами.

Таким образом, данное исследование имеет практическое значение, поскольку позволяет определять характеристики композитных материалов с помощью моделирования, что способствует созданию новых и улучшению существующих промышленных композитных материалов с желаемыми свойствами.

В данной аналитической записке будет рассмотрен разведочный анализ данных и проведен подбор гиперпараметров для нескольких моделей машинного обучения, таких как Линейные модели, Деревья решений, Ансамблирование и Модели ближайших соседей. Также будет оценена базовая модель для сравнения с другими моделями. В рамках работы будет проведено обучение нейронной сети MLPRegressor и Sequential.

В рамках данного исследования было разработано приложение в виде командной строки, которое позволяет получать прогнозы характеристик композитных материалов на основе алгоритм машинного обучения Случайный лес.

Также был создан репозиторий в GitHub, в котором размещен код исследования. Данный подход к разработке приложения и оформлению кода позволяет другим специалистам в области композитных материалов использовать наши результаты и модели для создания новых материалов с оптимальными свойствами.

# **Аналитическая часть**

# **Постановка задачи**

Цель данной работы заключается в разработке модели прогнозирования конечных свойств новых композиционных материалов имея данные о начальных свойствах исходных компонентов.

Результаты данной работы будут использоваться для уменьшения количества физических испытаний, что позволит сократить затраты на производство и ускорить процесс разработки новых материалов. Кроме того, созданные модели могут быть использованы для пополнения базы данных материалов и создания цифровых двойников новых композитов.

В первую очередь необходимо провести разведочный анализ данных. Знакомство с данными необходимо для моделирования в машинном обучении. На этом этапе происходит импорт и предварительный просмотр данных, выявление основных характеристик, таких как количество переменных, объем выборки, типы данных и наличие пропусков. Также производится очистка данных от выбросов, преобразование данных (нормализация, стандартизация или закодирование категориальных данных) и извлечение наиболее значимых признаков для использования в модели машинного обучения. Наконец, выборка разделяется на обучающую и тестовую. В данной работе используется 70 процентов данных для обучения и 30 процентов для тестирования модели.

# **Описание используемых методов**

Данная задача относится к машинному обучению с учителем и традиционно решается как задача регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её, поэтому для наилучшего решения в процессе исследования были применены следующие методы:

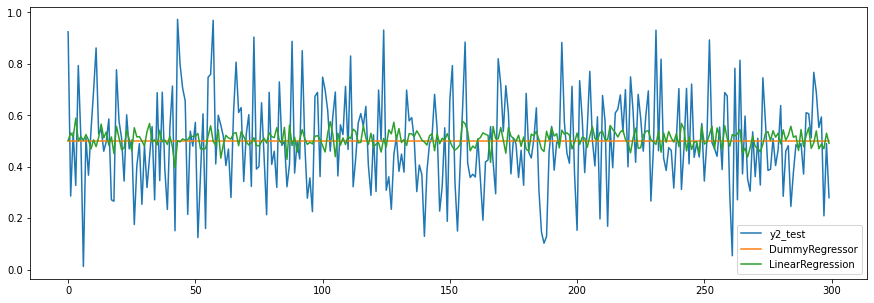
1. линейные модели такие как линейная регрессия, Лассо и метод опорных векторов;
2. ансамблевые алгоритмы машинного обучения с учителем на основе деревьев решений: дерево решений и случайный лес;
3. метод оптимизации на основе градиентного спуска: стохастический градиентный спуск;
4. модели ближайших соседей: в данном исследовании использовалась модель к-ближайших соседей.
5. базовые модели: модель фиктивного регрессора. Базовые модели используются для сравнения с более сложными моделями и оценки их качества.
6. нейронные сети: в данном исследовании использовался многослойный перцептрон и Sequential

Рассмотри преимущества и недостатки каждого метода более подробно.

Линейные модели - это методы анализа данных, которые используются для построения прогнозных моделей, основанных на линейной комбинации независимых переменных.

Линейная регрессия является одним из самых простых и наиболее используемых методов прогнозирования. Он использует линейную комбинацию независимых переменных для прогнозирования зависимой переменной. Преимущества линейной регрессии является простота реализации, хорошую интерпретируемость и устойчивость к шуму в данных. Недостатком линейной регрессии является то, что она не может моделировать нелинейные зависимости между переменными.

Рисунок 1 – График Линейной регрессии для прогнозирования прочности при растяжении.



На рисунке 1 изображена линейная регрессия для прогнозирования прочности при растяжении. Коэффициент детерминации имеет значение -0,025 (Таблица 1), указывая на плохую прогнозную способность даже по сравнению с базовой моделью. Средняя абсолютная ошибки также отрицательна (-0,168 и -0,133 соответственно), что говорит о несоответствии знаков фактических и прогнозных значений. Максимальная ошибка равна -0,442, что указывает на большое отклонение между наблюдаемыми и предсказанными данными. Таким образом, модель не способна адекватно описать зависимость между переменными и не может быть использована для прогнозирования.

Лассо - это метод регрессии, которая учитывает только наиболее важные независимые переменные. Одним из преимуществ Лассо является то, что он уменьшает количество независимых переменных, что может уменьшить переобучение. Недостатком Лассо является то, что он неспособен уловить нелинейные зависимости между переменными. В таблице 1 отображены метрики по модели с использованием метода Лассо. Результаты хуже базовой модели.

Метод опорных векторов (SVM) - это метод классификации, который разделяет данные на две категории с помощью разделяющей гиперплоскости, которая максимизирует расстояние между точками и гиперплоскостью. Преимуществом метода является то, что он может обрабатывать как линейные, так и нелинейные зависимости между переменными. Недостатком является его сложность реализации и необходимость настройки гиперпараметров. Метрики модели SVM также неудовлетворительные.

Таблица 1 – Метрики оценки качества линейных моделей для прогнозирования прочности при растяжении

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | Среднеквадратичная ошибка | Средняя абсолютная ошибка | Максимальная ошибка |
| Фиктивный регрессор | -0,011 | -0,167 | -0,132 | -0,444 |
| Лассо | -0,011 | -0,167 | -0,132 | -0,444 |
| Линейная регрессия | -0,025 | -0,168 | -0,133 | -0,442 |

Коэффициент детерминации R2 измеряет, насколько хорошо модель соответствует данным. Значение коэффициента может быть от 0 до 1, где 1 означает, что модель идеально соответствует данным, а 0 означает, что модель не объясняет никакой вариации в данных.

Отрицательное значение коэффициента детерминации означает, что модель плохо соответствует данным и не объясняет вариации. Это может быть связано с плохой корреляцией между предсказываемой переменной и объясняющими переменными. Если корреляция между этими переменными низкая, то модель сложно предсказать значения предсказываемой переменной на основе объясняющих переменных.

В вашем примере все три линейные модели имеют отрицательные значения коэффициент детерминации, что указывает на то, что они плохо соответствуют данным. Это подтверждается и другими метриками: такими как среднеквадратичная ошибка, средняя абсолютная ошибка, максимальная ошибка, они тоже являются отрицательными. Это означает, что результаты хуже, чем если бы мы просто использовали константу (DummyRegressor).

Деревья решений являются относительно простыми и интерпретируемыми моделями, что делает их полезными для понимания причинно-следственных связей в данных. Преимущества модели заключается в том, что они могут работать с различными типами признаков и могут использоваться как для задач классификации, так и для задач регрессии. Недостатки данной модели являются склонность к переобучению, особенно при работе с большим количеством признаков и глубокими деревьями.

Таблица 2 – Метрики ансамблевых алгоритмов на основе деревьев решений для прогнозирования прочности при растяжении

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | Среднеквадратичная ошибка | Средняя абсолютная ошибка | Максимальная ошибка |
| Случайный лес | 0,004 | -0,161 | -0,131 | -0,443 |
| Деревья решений | -0,008 | -0,167 | -0,133 | -0,441 |

В таблице два представлены метрики по модели Случайный лес и Деревья решений. Обе модели не являются достаточно точными для предсказания значений целевой переменной, но у Случайного леса наблюдается уже положительный коэффициент детерминации.

Случайный лес является ансамблем деревьев решений, который может повысить точность предсказаний и уменьшить склонность к переобучению по сравнению с одиночными деревьями решений. Кроме того, случайный лес может автоматически выбирать признаки для каждого дерева, что позволяет сократить время обучения и уменьшить вероятность переобучения. Недостаток случайного леса является сложность и меньшая интерпретированность, чем отдельные деревья решений, а также может требовать больших объемов данных для достижения лучшей производительности.

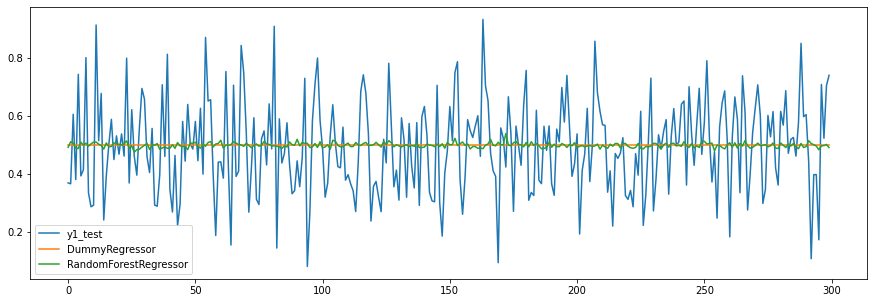
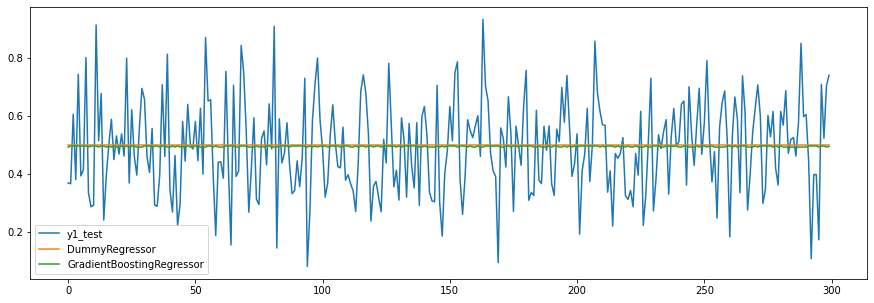


Рисунок 2 – График Случайного леса для прогнозирования модуля упругости при растяжении

Стохастический градиентный спуск (GBR) является алгоритмом машинного обучения, который используется для решения задач регрессии. Он основан на ансамбле деревьев решений и позволяет достигнуть высокой точности прогнозирования.

Одним из главных преимуществ этого метода является его способность обрабатывать различные типы данных, такие как числовые, текстовые, категориальные. Это делает его универсальным инструментом для работы с данными различной природы. Алгоритм также позволяет работать с нелинейными зависимостями между переменными, что расширяет его возможности.

Недостатком данной модели является склонность к переобучению, если не установить оптимальные значения гиперпараметров. Кроме того, алгоритм может требовать большое количество времени и вычислительных ресурсов для обучения модели, особенно при работе с большими объемами данных. На данной задаче Градиентный спуск показал плохие результаты, R2 равен -0,012

 Рисунок 3 – График Стохастический градиентный спуск для прогнозирования модуля упругости при растяжении

К-ближайщих соседей - это алгоритм машинного обучения, который используется для регрессии на основе метода ближайших соседей. Одним из главных преимуществ методов является его простота в использовании. Он не требует настройки сложных параметров модели, что делает его хорошим выбором для быстрого и простого решения задач регрессии. Кроме того, он хорошо работает с нелинейными зависимостями между признаками и целевой переменной.

Таблица 2 – Метрики оценки качества моделей после подбора гиперпараметров для прогнозирования модуля упругости при растяжении.

Однако метод К-ближайщих соседей чувствителен к выбросам, потому что признает все точки в качестве соседей, что может приводить к ошибочным прогнозам. Кроме того, он может быть вычислительно затратным для больших наборов данных и может требовать большого объема памяти для хранения всего обучающего набора.

Другой недостаток метода связан с тем, что он не может захватывать глобальные зависимости между признаками и целевой переменной. Он может обнаруживать только локальные зависимости, поэтому если зависимости глобальные, то алгоритм может давать плохие результаты. И наконец, метод ближайших соседей может иметь проблемы с мультиколлинеарностью, когда признаки сильно коррелируют друг с другом. В таких случаях он может давать неправильные прогнозы и требовать более сложных методов обработки признаков, таких как снижение размерности.

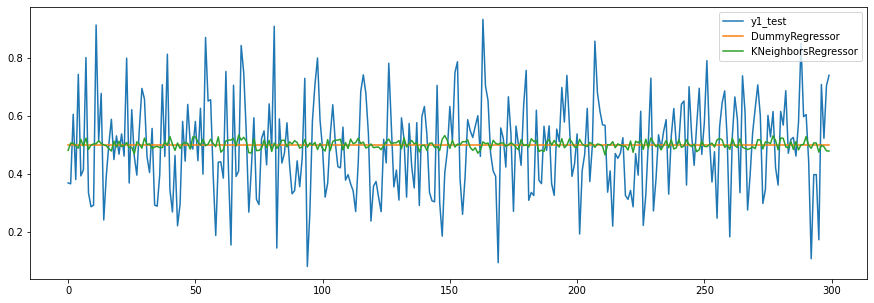


Рисунок 4 – График К-ближайщих соседей для прогнозирования модуля упругости при растяжении

После подбора гиперпараметров модель показала положительный коэффициент детерминации, он составил 0.003287, что означает, что модель плохо объясняет результаты, но результаты уже лучше, чем у базовой модели.

При это среднеквадратичная ошибка равна -0.161188, что необычно, так как MSE не может быть отрицательным числом. Возможно необходима дополнительная предобработка данных с экспертом.

Многослойный перцептрон (MLP) - это нейронная сеть, которая состоит из нескольких слоев нейронов. Каждый нейрон в одном слое связан со всеми нейронами в предыдущем и следующем слоях. MLP является одним из самых распространенных типов нейронных сетей и широко используется в различных приложениях машинного обучения, включая распознавание образов, классификацию, регрессию и прогнозирование.

Преимущества MLP включают гибкость, возможность обработки сложных данных и распараллеливание. MLP может быть настроен для решения широкого спектра задач машинного обучения, используя различные функции активации и алгоритмы обучения. Он может обрабатывать данные с высокой степенью сложности и нелинейности, моделируя нелинейные отношения между входными и выходными данными. MLP также может эффективно использовать многопроцессорные и многоядерные системы, что позволяет ускорить процесс обучения и улучшить производительность.

Недостатки MLP включают необходимость большого количества данных для обучения, склонность к переобучению и вычислительную сложность. MLP требует большого объема данных для эффективного обучения, особенно при большом числе скрытых слоев. Если количество слоев и нейронов слишком велико, а данных недостаточно, MLP может быть склонен к переобучению. Обучение MLP может быть вычислительно сложным и требовать значительных вычислительных ресурсов, особенно для больших сетей и большого объема данных.

Таблица 3 – Метрики оценки качества многослойного перцептрона для прогнозирования соотношения матрица-наполнитель

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | Среднеквадратичная ошибка | Средняя абсолютная ошибка | Максимальная ошибка |
| Фиктивный регрессор | -0,001 | -0,179 | -0,144 | -0,460 |
| Многослойный перцептрон | -0,003 | -0,179 | -0,143 | -0,461 |

В данном конкретном случае, результаты многослойного перцептрона оказались хуже, чем у базовой модели, что говорит о неэффективности данной архитектуры нейронной сети для данной конкретной задачи.

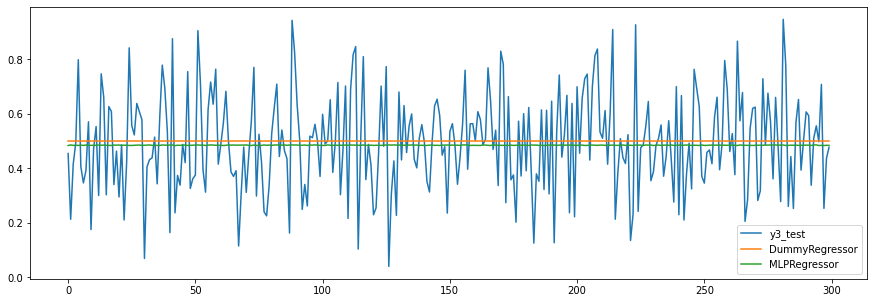


Рисунок 5 – График Многослойный перцептрон для прогнозирования соотношения матрица-наполнитель

Sequential - это одна из моделей нейронных сетей в библиотеке Keras, которая представляет собой последовательность слоев. Он позволяет создавать модели путем добавления слоев друг за другом в последовательном порядке. Это делает создание моделей нейронных сетей быстрым и простым.

Стоит отметить, что Sequential имеет свои ограничения и недостатки. Например, Sequential не является наилучшим выбором для создания сложных моделей, которые включают в себя несколько входных или выходных слоев. Кроме того, Sequential не предоставляет полного контроля над процессом оптимизации, что может быть важным при работе с некоторыми типами данных.

Таблица 4 – Метрики оценки качества Sequential для прогнозирования соотношения матрица-наполнитель

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Коэффициент детерминации | Среднеквадратичная ошибка | Средняя абсолютная ошибка | Максимальная ошибка |
| Фиктивный регрессор | -0,001 | -0,179 | -0,144 | -0,460 |
| Sequentail | -0,308 | -0,205 | -0,163 | -0,570 |
| Sequential + EarlyStopping | -0,019 | -0,181 | -0,146 | -0,465 |
| Sequential + Droupout | -0,314 | -0,205 | -0,165 | -0,536 |

В таблице отображено, что лучший коэффициент детерминации R2 для модели Sequential + EarlyStopping равен -0.018821, что означает, что модель не объясняет вариацию целевой переменной и выдает результаты хуже, чем случайное угадывание.

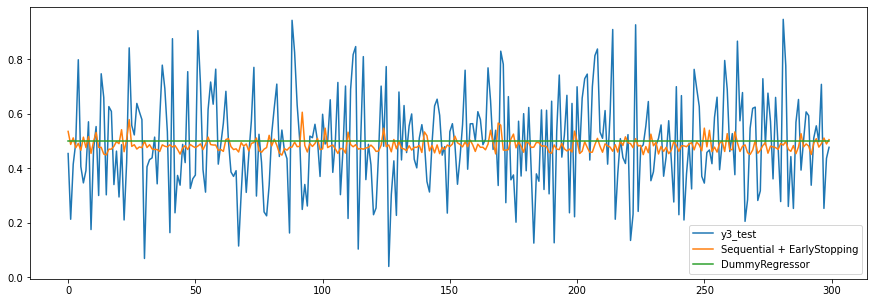


Рисунок 6 – График Sequential + EarlyStopping для прогнозирования соотношения матрица-наполнитель

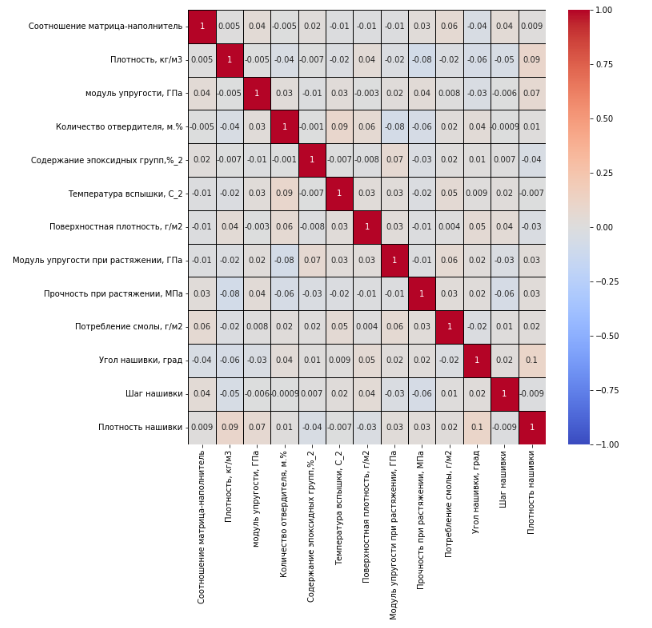
## **Разведочный анализ данных**

Предварительная обработка данных является важным этапом в анализе данных и моделировании машинного обучения, который позволяет обеспечить правильность, полноту и достоверность данных для использования в моделях машинного обучения и принятия на их основе обоснованных решений.. Она включает в себя несколько этапов, таких как:

Она включает в себя несколько этапов:

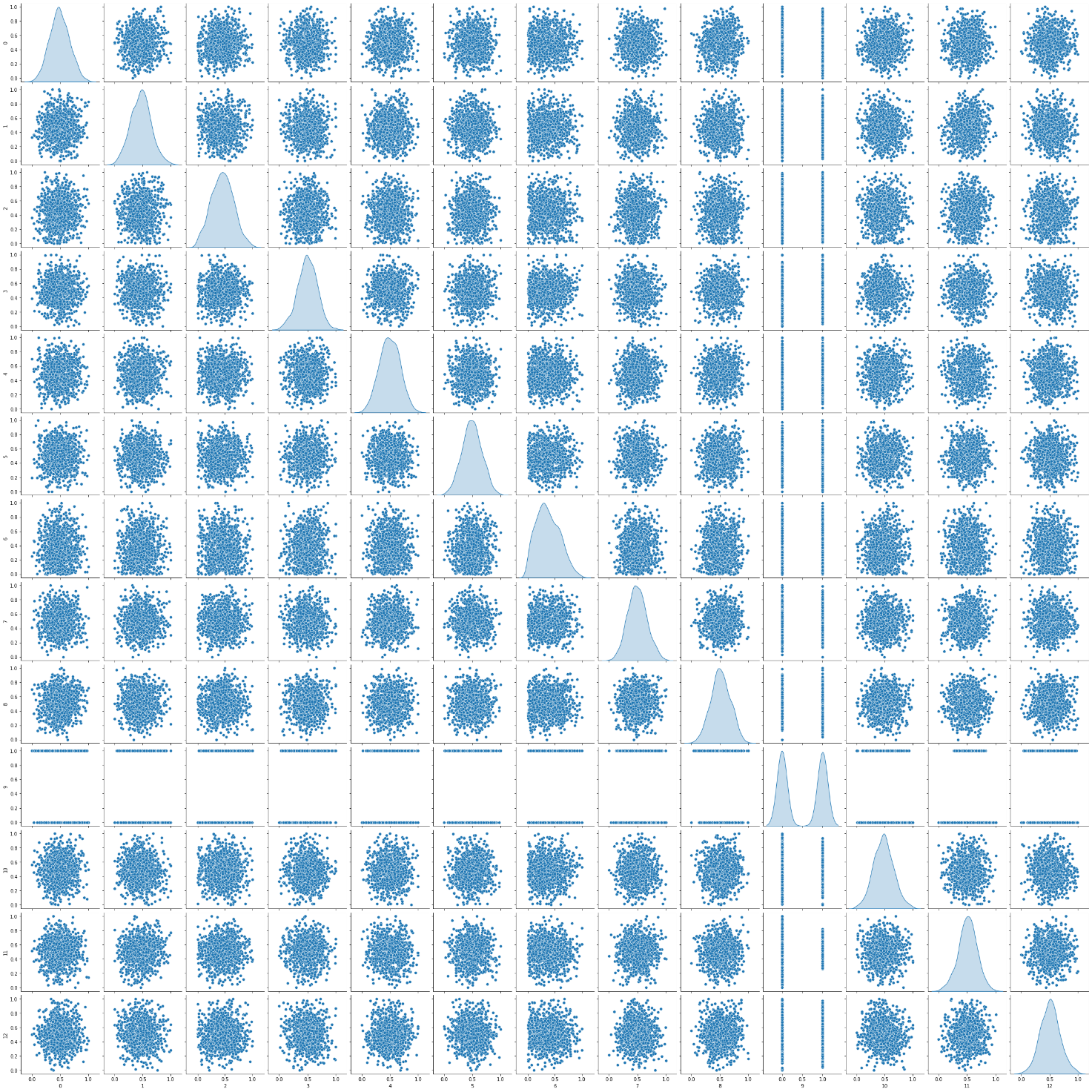
1. импорт и предварительный просмотр данных;
2. выявление основных характеристик данных: количество переменных, объем выборки, типы данных, наличие пропусков;
3. очистка данных: нахождение выбросов и удаления их из выборки или замена их средними значениями;
4. преобразование данных: нормализация, стандартизация или закодирование категориальных данных для унификации масштаба и формата;
5. извлечение признаков: отбор наиболее значимых признаков для использования в модели машинного обучения с помощью корреляционного анализа или методов понижения размерности;
6. разделение выборки на обучающую и тестовую, с использованием 70% данных для обучения и 30% для тестирования модели и оценки ее точности.

На этапе ознакомления с данными было выявлено, что объединенный датасет содержит 13 переменных, из которых 12 являются входными и одна переменная является выходной. Объем выборки равен 1023 строкам, что является достаточным для обучения модели машинного обучения с учителем. Все столбцы имеют тип данных float64, кроме столбца Угол нашивки, который имеет тип данных int64. Это означает, что все переменные являются числовыми и не требуют кодирования категориальных данных. Пропуски в данных отсутствуют. Это позволяет использовать полный объем выборки для анализа и построения модели.

Рисунок 7 – Тепловая карта корреляционной матрицы

В рамках разведочного анализа данных была построена тепловая карта корреляционной матрицы, она изображена на рисунке 7. Тепловая карта - это графическое представление коэффициентов корреляции Пирсона между всеми парами переменных в датасете. Оно может принимать значения от -1 до 1. Чем ближе коэффициент корреляции к 1 или -1, тем сильнее связь, а чем ближе к 0, тем слабее связь. Тепловая карта корреляционной матрицы помогает визуализировать корреляцию между всеми парами переменных в датасете и выявлять возможные проблемы мультиколлинеарности. Мультиколлинеарность - это ситуация, когда две или более переменных сильно коррелируют друг с другом, что может приводить к нестабильности и низкой точности модели машинного обучения. По визуализации видно, что тепловая карта корреляционной матрицы датасета имеет едва заметный оттенок, что свидетельствует о минимальной корреляции между переменными в датасете.

Также был построен Pairplot (рисунок 8), также известный как график попарных отношений. График представляет из себя является матрицу графиков, каждый из которых показывает связь между каждой парой переменных в датасете. Этот тип графика используется для наглядной демонстрации корреляций, распределений и нелинейных зависимостей в данных.

 Рисунок 8 - График попарных отношений

В процессе очистки данных был использован метод z-оценки. На рисунке 9 изображены диаграммы под название ящик с усами, выбросу обозначены красным цветом.

Z-оценка представляет собой меру того, насколько значение отклоняется от среднего значения в единицах стандартного отклонения. Обычно выбросы определяются как значения, которые находятся за пределами определенного количества стандартных отклонений от среднего значения. В данной работе был использован порог в 3 стандартных отклонения. Больше всего было обнаружено выбросов в переменной "Плотность нашивки", а в остальных переменных было небольшое количество выбросов.

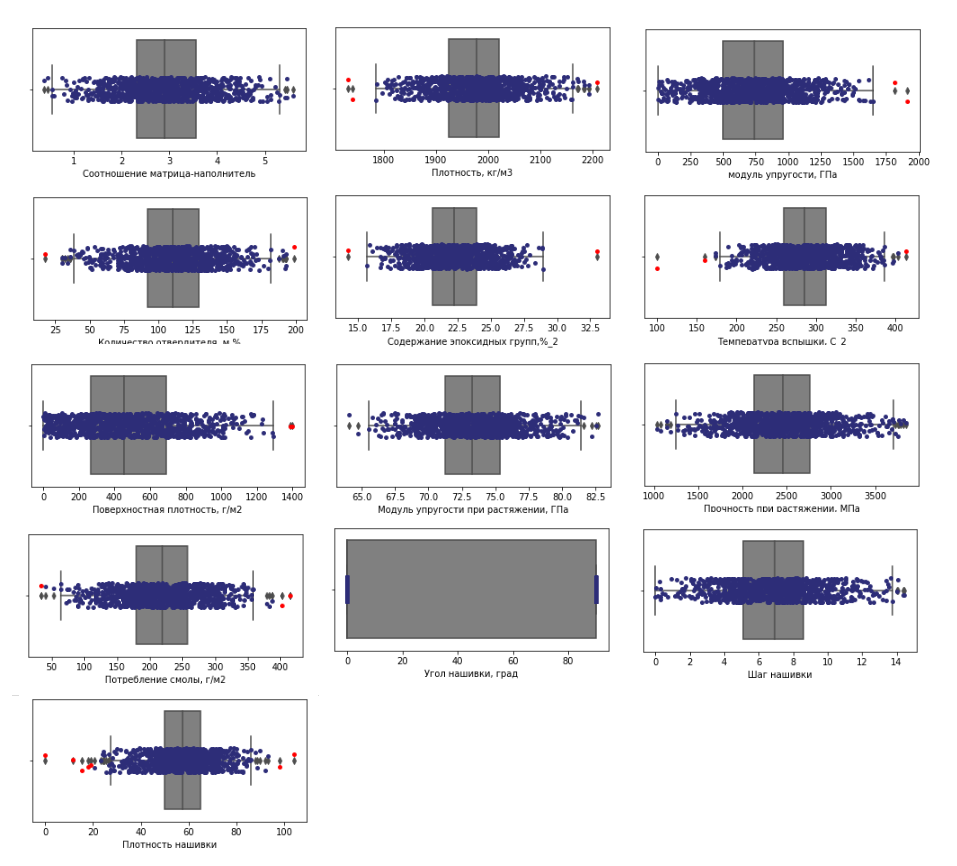
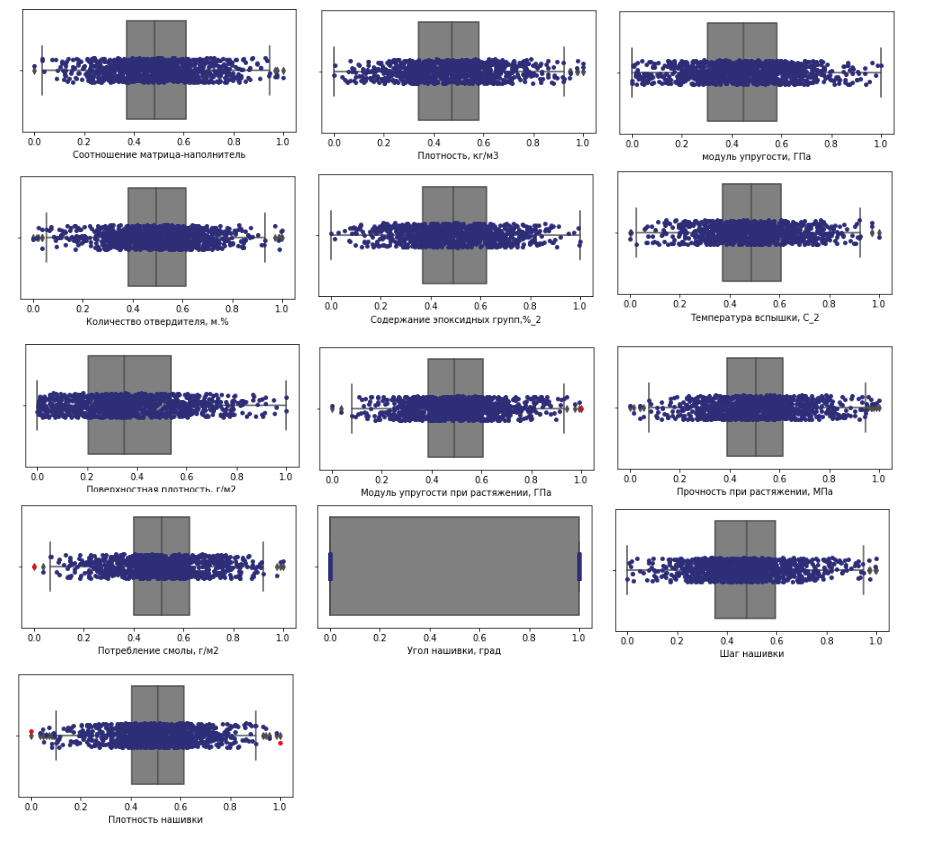


Рисунок 9 – Диаграмма Ящик с усами до удаления выбросов

 Рисунок 10 – Диаграмма ящик с усами после удаления выбросов

Следующей этап – это преобразование данных.

Этот этап может включать в себя нормализацию, стандартизацию или закодирование категориальных данных методом "one-hot encoding". В данной работе будет рассмотрен метод нормализации.

Нормализация данных - это процесс масштабирования признаков, чтобы они находились в определенном диапазоне. Один из наиболее распространенных методов нормализации данных - это MinMaxScaler. Он приводит значения признаков к диапазону от 0 до 1.

Применение MinMaxScaler имеет несколько преимуществ. Во-первых, он сохраняет форму распределения данных, что важно для некоторых алгоритмов машинного обучения. Во-вторых, он обеспечивает легкость интерпретации масштабированных значений, что облегчает анализ результатов. В-третьих, он уменьшает влияние выбросов на модель, поскольку они также масштабируются в пределах диапазона. Наконец, он улучшает сходимость алгоритмов машинного обучения, поскольку признаки находятся в более узком диапазоне значений. На рисунке 11 отображен график плотности распределения до нормализации. График показывает, как часто встречаются разные значения переменной в выборке или в совокупности. Он сглаживает гистограмму и приближает форму теоретического распределения. Это позволяет сравнивать распределения разных переменных на одном графике.

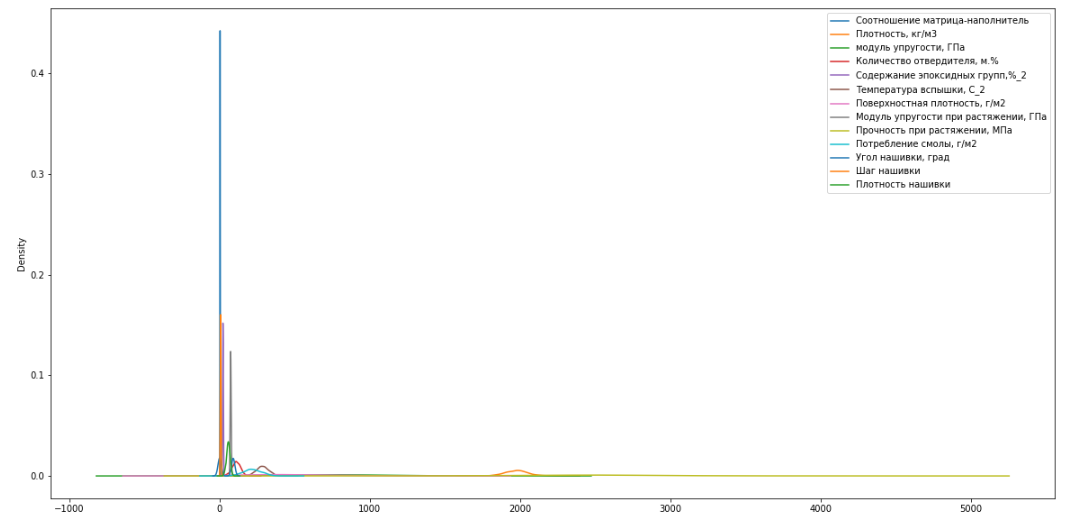


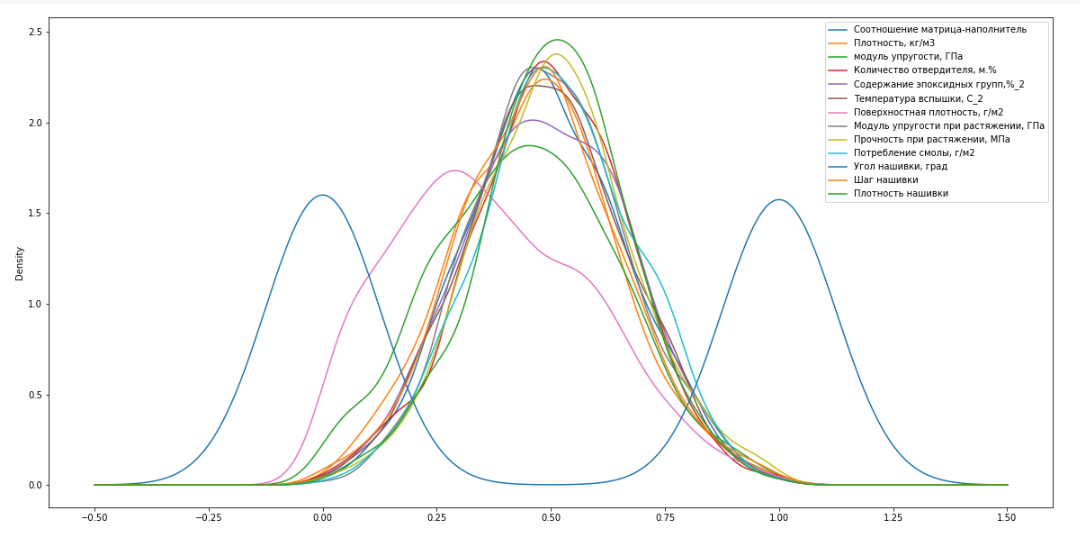
Рисунок 11 – График плотности распределения до нормализации

Рисунок 12 - График плотности распределения после нормализации

После нормализации данных статистические характеристики изменились. В частности, средние значения признаков стали ближе к нулю, а стандартные отклонения признаков сократились. Также можно заметить, что все признаки теперь имеют примерно одинаковый масштаб, что позволяет более корректно сравнивать их между собой (рисунок 12). Переменная "Угол нашивки" преобразована до значения 0 или 1.

Нормализация данных является важным шагом при подготовке данных для машинного обучения, так как это помогает улучшить производительность моделей и улучшить качество их прогнозов.

На этапе извлечения признаков из данных происходит отбор наиболее значимых признаков для использования в модели машинного обучения. Для построения модели в данной работе было решено оставить 30% данных на тестирование, в то время как на оставшихся 70% данных происходит обучение моделей. Такой подход позволяет оценить качество модели на новых данных, не задействованных в обучении, и протестировать ее на способность обобщения.

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

Описательная статистика позволяет извлекать полезную информацию из данных и предоставлять ее в виде числовых показателей, таких как среднее значение, медиана, дисперсия и стандартное отклонение (таблица 5). Стандартное отклонение показывает степень разброса данных относительно среднего значения. Чем больше стандартное отклонение, тем больше диапазон изменения признака в наборе данных. Например, стандартное отклонение Прочность при растяжении равно 485.6 ГПа, что говорит о том, что прочность растяжении сильно различается для разных образцов материала.

Минимальное и максимальное значение помогают определить наличие выбросов или аномальных значений в данных. Для данных по признаку Угол нашивки наблюдается только два варианта значений 0 градусов и 90 градусов.

В данном датасете наблюдается большой разброс по минимальному и максимальному значению в признаках по столбцу "Прочность при растяжении", "Поверхностная плотность", "Модуль упругости". Значения в данных столбцах может значительно меняться в зависимости от других факторов.

Квартили показывают распределение данных. Они помогают определить форму распределения и наличие асимметрии или скошенности в данных. Большинство значений распределены равномерно, это говорит о преобладании нормального распределения в данных.

Таблица 5 – Описательная статистика

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| Соотношение матрица-наполнитель | 1000,00 | 2,94 | 0,91 | 0,39 | 2,32 | 2,91 | 3,55 | 5,59 |
| Плотность, кг/м3 | 1000,00 | 1975,40 | 72,95 | 1784,48 | 1923,63 | 1977,32 | 2021,16 | 2192,74 |
| модуль упругости, ГПа | 1000,00 | 738,68 | 327,55 | 2,44 | 500,77 | 741,15 | 961,65 | 1649,42 |
| Количество отвердителя, м.% | 1000,00 | 110,82 | 27,87 | 29,96 | 92,52 | 110,65 | 129,85 | 192,85 |
| Содержание эпоксидных групп,%\_2 | 1000,00 | 22,24 | 2,38 | 15,70 | 20,58 | 22,22 | 23,97 | 28,96 |
| Температура вспышки, С\_2 | 1000,00 | 285,96 | 40,23 | 173,48 | 259,10 | 285,85 | 313,03 | 403,65 |
| Поверхностная плотность, г/м2 | 1000,00 | 479,86 | 277,71 | 0,60 | 266,98 | 450,87 | 691,53 | 1291,34 |
| Модуль упругости при растяжении, ГПа | 1000,00 | 73,32 | 3,11 | 64,05 | 71,25 | 73,23 | 75,33 | 82,68 |
| Прочность при растяжении, МПа | 1000,00 | 2464,86 | 485,02 | 1036,86 | 2134,54 | 2456,39 | 2760,16 | 3848,44 |
| Потребление смолы, г/м2 | 1000,00 | 218,25 | 58,94 | 41,05 | 179,81 | 218,70 | 257,47 | 386,90 |
| Угол нашивки, град | 1000,00 | 44,64 | 45,02 | 0,00 | 0,00 | 0,00 | 90,00 | 90,00 |
| Шаг нашивки | 1000,00 | 6,91 | 2,56 | 0,04 | 5,11 | 6,92 | 8,59 | 14,44 |
| Плотность нашивки | 1000,00 | 57,28 | 11,85 | 20,57 | 49,89 | 57,47 | 64,93 | 92,96 |

После изучения описательной статистики был применен метод z-оценка, которая позволяет определить выбросы как значения, за пределами 3 стандартных отклонений от среднего значения. Для удобства отображения выбросы были подсвечены красным цветом на диаграмме ящик с усами (рисунок 13).

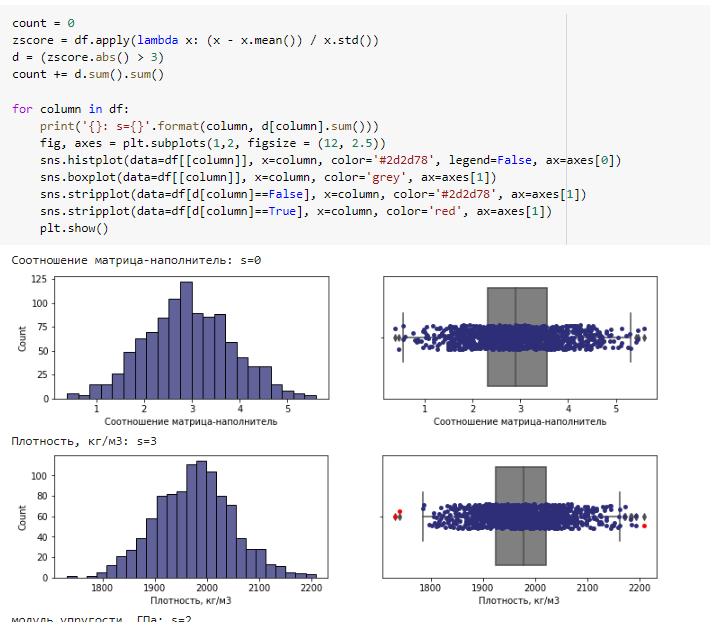


Рисунок 13 –Метод z-оценки

Так как данные имеют большой разброс была произведена нормализация данных с помощью метода MinMaxScaler. Это позволило привести данные к единому масштабу и увеличить эффективность алгоритмов машинного обучения и оптимизации. После нормализации данные находятся в диапазоне от 0 до 1, при этом распределения данных не изменилось (рисунок 13).

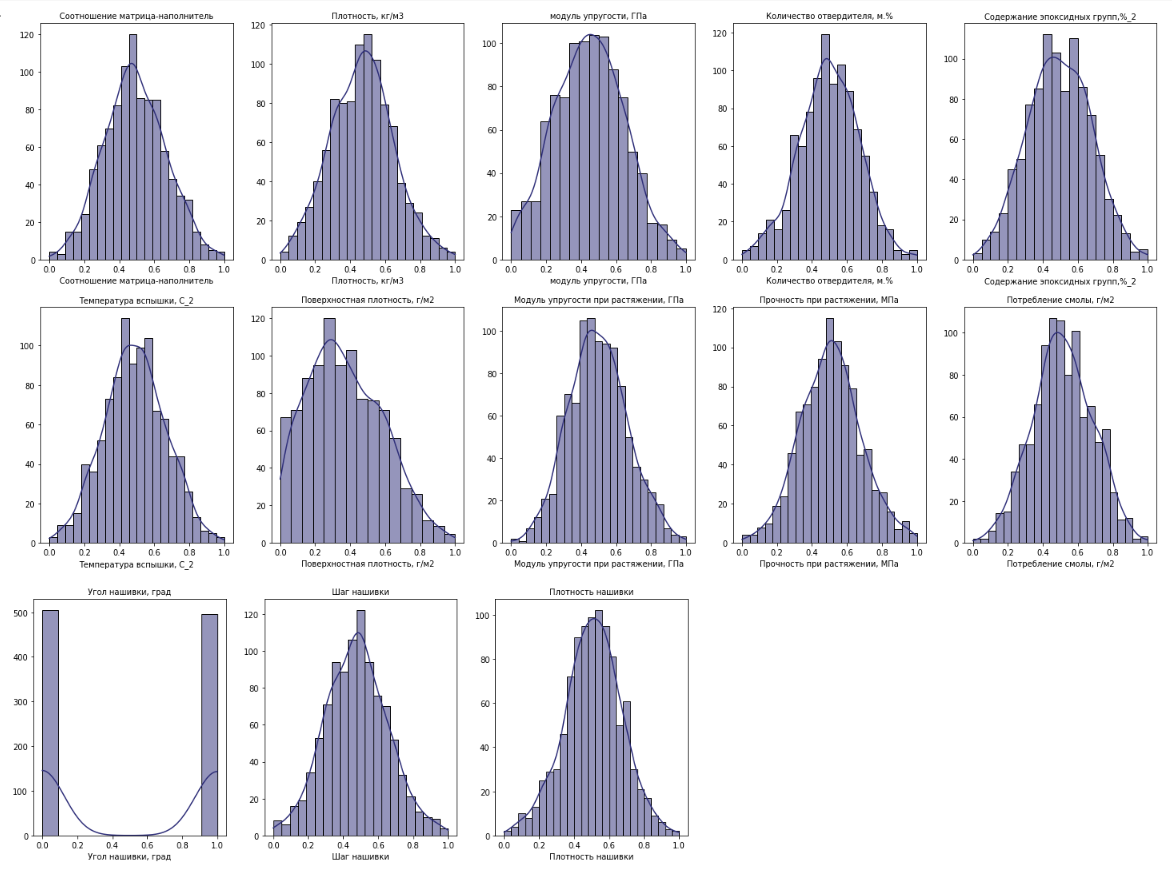


Рисунок 13 - График распределения после нормализации данных

данных, которая позволила.

## **Разработка и обучение модели**

Для прогноза модуля упругости при растяжении и прочности при растяжении было решено использовать следующие модели:

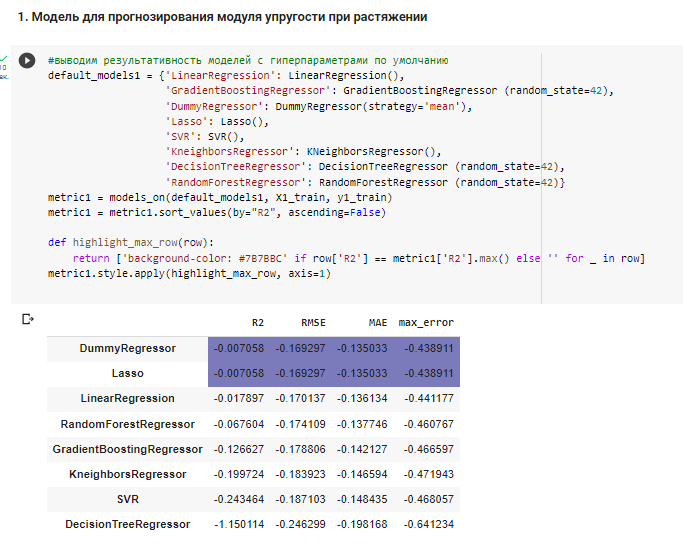
1) линейные модели: Линейная регрессия, Лассо и Метод опорных векторов.;

2) деревья решений и Случайный лес;

3) стохастический градиентный спуск.

4) модель К-ближайщих соседей.

5) базовые модели: в данном исследовании использовалась модель фиктивный регрессор. Базовые модели используются для сравнения с более сложными моделями и оценки их качества.

 Рисунок 14 – Оценка качества моделей для прогнозирования модуля упругости при растяжении

В данном исследовании был произведен поиск гиперпараметров для каждой модели с помощью поиска по сетке с перекрестной проверкой. Данный подход позволяет выбрать наилучшие значения гиперпараметров для модели.

В данной работе был использован поиск по сетке с перекрестной проверкой для поиска оптимальных значений гиперпараметров. Поиск по сетке состоит из следующих шагов: определение набора возможных значений для каждого гиперпараметра, обучение модели со всеми возможными комбинациями этих значений и оценка качества модели с помощью перекрестной проверки.

Перекрестная проверка – это метод оценки качества модели, который используется для измерения производительности модели на неизвестных данных. Данный метод предполагает разбиение исходных данных на 10 блоков (или “складок”), обучение модели на 9 блоках и тестирование на оставшемся блоке. Этот процесс повторяется 10 раз, так что каждый блок используется для тестирования один раз. Затем результаты усредняются для всех 10 разбиений и так получается оценка качества модели для данной комбинации гиперпараметров.

Таким образом, использование поиска по сетке с перекрестной проверкой является эффективным методом для выбора оптимальных значений гиперпараметров модели машинного обучения. Данный подход позволяет улучшить качество модели и повысить ее точность прогнозирования.

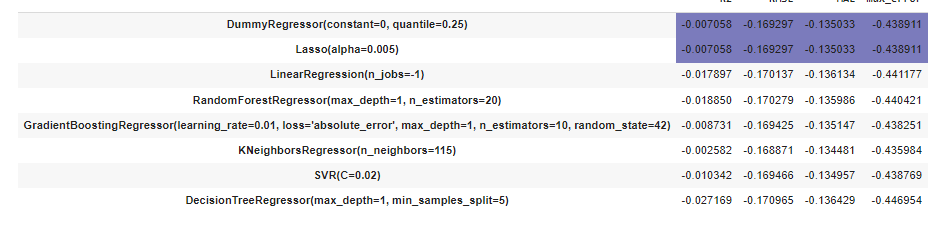


Рисунок 14 – Оценка качества моделей для прогнозирования модуля упругости при растяжении после поиска гиперпараметров

После проведения поиска гиперпараметров для разных моделей машинного обучения были сравнены их метрики качества на тестовых данных. Было обнаружено, что подбор гиперпараметров улучшил результаты моделей по сравнению с их значениями по умолчанию (рисунок 14).

Для прогнозирования модуля упругости при растяжении были выбраны три модели: Случайный лес, метод к-ближайщих соседей и стохастический градиентный спуск. Эти модели имеют более высокий коэффициент детерминации R2, который показывает, насколько хорошо модель объясняет вариацию данных, и наименьшие значения RMSE, MAE и max\_error, которые измеряют ошибку прогноза модели.

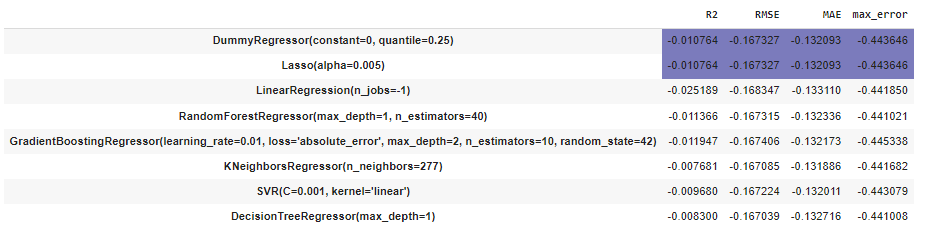
Качество моделей для для прогнозирования прочности при растяжении хуже, чем для прогноза для прогнозирования модуля упругости при растяжении (рисунок 15). Это может быть связано с тем, что прочность при растяжении более чувствительна к различным факторам, таким как дефекты материала или условия нагружения. Воспользуемся такими же регрессорами как и для первого прогноза. Изменим только Градиентный спуск на линейную регрессию.

Рисунок 15 – Оценка качества моделей для прогнозирования прочности при растяжении поиска гиперпараметров

## **Тестирование модели**

Для прогнозирования модуля упругости при растяжении на тестовых данных были выбраны четырех регрессионных моделей: K-ближайщих соседей, Случайный лес, Стохастический градиентный спуск и Базовая модель для сравнения.

Для оценки качества прогнозов были вычислены метрики, включая среднеквадратическую ошибку, среднюю абсолютную ошибку, коэффициент детерминации и максимальную ошибку (рисунок 16).

В результате сравнения моделей было выяснено, что лучшими моделями является Случайный лес и К-ближайщих соседей, они имеют наибольшее значение коэффициента детерминации R2, но так как это значение все равно близкое к нулю это значит, что модели объясняет небольшую часть изменчивости данных. Эти модель можно использовать для дальнейшего прогнозирования с помощью приложения.

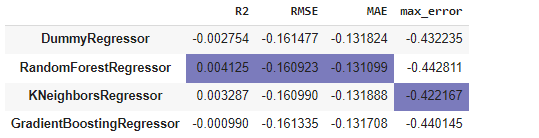


Рисунок 16 – Оценка качества моделей для прогнозирования модуля упругости при растяжении после обучения

Для прогнозирования прочности при растяжении были использованы 4 модели: Базовая модель для сравнения, Случайный лес, К-ближайщих соседей, Линейная регрессия. В соответствии с данными R2, RMSE, MAE и max\_error, лучшей моделью является DummyRegressor. Однако, все модели имеют отрицательные значения R2, это значит, что модели не могут объяснить вариацию в данных (рисунок 17).

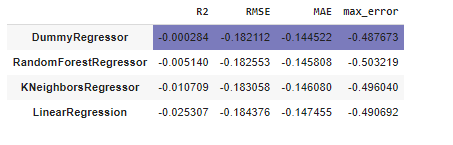


Рисунок 17 – Оценка качества моделей для прогнозирования прочности при растяжении после обучения

## **Написать нейронную сеть, которая будет рекомендовать соотношение матрица.**

Для прогнозирования соотношение матрицы-наполнителя в композитных материалах была использована модель Многослойный персептрона. Это тип искусственной нейронной сети, которая состоит из нескольких слоев нейронов, соединенных взвешенными связями. Многослойный персептрон может выполнять задачи регрессии, то есть предсказывать непрерывные значения на основе входных данных, а также обрабатывать нелинейные зависимости между входными и выходными данными и работать с данными, которые не являются линейно разделимыми.

На рисунке 18 изображен график Кривая потерь, с помощью этого графика можноввизуализировать изменения функции потерь на каждой итерации обучения модели MLPRegressor, что помогает оценить эффективность процесса обучения и выявить возможные проблемы.

На рисунке видно, что после 20-й итерации обучения модель переобучилась и достигла минимума функции потерь, что означает, что она не может улучшить свои предсказания дальше.

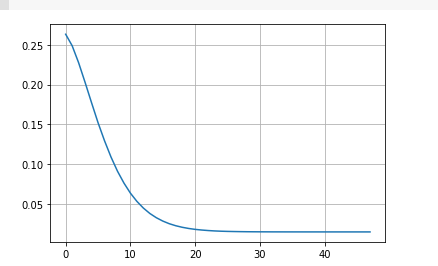


Рисунок 18 – График кривая потерь при обучении Многослойного персептрона

После тестирования на тестовых данных, коэффициент детерминации MLPRegressor оказался отрицательным это означает, что модель работает хуже, чем простая константная модель. Это может быть связано с тем, что модель не учитывает некоторые важные факторы или что данные не подходят для данной модели.

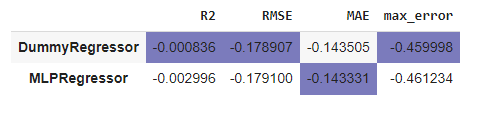


Рисунок 19 – Оценка качества моделей для прогнозирования прочности при растяжении после обучения

В рамках исследования для прогнозирования соотношения матрицы-наполнителя в композитных материалах была также применена модель нейронной сети Sequential. Это тип модели из библиотеки Keras, который позволяет создавать модели машинного обучения с помощью последовательности слоев. Слой - это базовый блок модели, который выполняет некоторую операцию над входными данными и выдает выходные данные. Каждый слой имеет один вход и один выход тензора. Тензор - это многомерный массив чисел, который хранит данные. Модель Sequential удобна для создания простых моделей, которые имеют линейную структуру, то есть данные проходят через слои один за другим без ветвлений или циклов.

Для модели Sequential были также использованы методы EarlyStopping и Dropout, которые помогают предотвратить переобучение модели. Переобучение означает, что модель слишком хорошо подстраивается под обучающие данные и теряет способность обобщать на новых данных. Метод EarlyStopping позволяет остановить обучение модели, когда ошибка на валидационных данных перестает уменьшаться или начинает увеличиваться. Это помогает избежать излишнего обучения модели и сохранить оптимальные веса модели (рисунок 20).

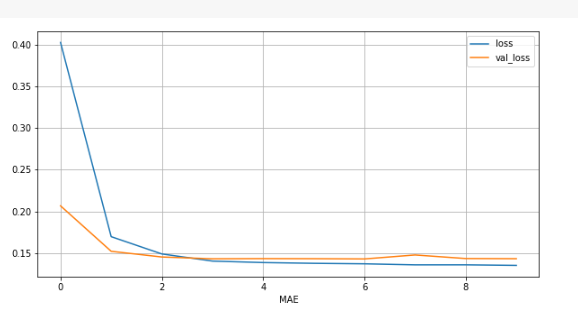


Рисунок 20 – Линейный график для ошибки MAE модель Sequential+ EarlyStopping

На графике видно, что ошибка MAE уменьшается с увеличением количества эпох обучения модели, но после 10-й эпохи она стабилизируется и не меняется существенно. Это означает, что модель достигла оптимального уровня обучения и не может улучшить свои предсказания дальше. Также видно, что ошибка MAE на валидационных данных немного выше, чем на обучающих данных, что свидетельствует о небольшом переобучении модели.

Метод Dropout позволяет случайным образом отключать некоторые нейроны в слое во время обучения модели. Это помогает уменьшить зависимость между нейронами и предотвратить переадаптацию модели к обучающим данным. Оба метода являются классами из библиотеки Keras, которые могут быть добавлены в модель Sequential как слои.

После проведения тестирования на тестовых данных и сбора метрик моделей в сводную таблицу, было обнаружено, что модель Sequential + EarlyStopping показала наилучшие результаты (рисунок 21).

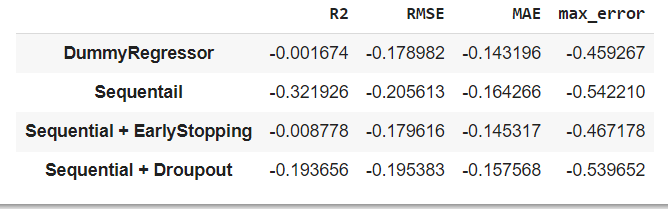
**

Рисунок 21 – Оценка качества моделей для прогнозирования прочности при растяжении после обучения

Значение R2 для этой модели было наибольшим, а значения RMSE, MAE и max\_error - наименьшими среди всех вариантов использования Sequential. Эти результаты указывают на то, что модель Sequential + EarlyStopping лучше обобщает данные и адаптируется к новым данным, чем Sequential со стандартными настройками или со слоями Dropout. Однако все значения метрик отрицательные, это свидетельствует о том, что модели работают хуже, чем простая константная модель. Это может быть связано с тем, что в данных есть некоторые скрытые факторы, которые не учитываются моделями.

## **Разработка приложения**

В ходе работ было разработано приложение с интерфейсом командной, которое позволяет пользователям быстро и эффективно решать задачи, связанные с прогнозированием конечных свойств новых материалов. Пользователь вводит значения различных переменных, таких как соотношение матрица-наполнитель, плотность, модуль упругости, и так далее.

После того как пользователь ввел все необходимые данные, они проходят через предварительную обработку, включая масштабирование переменных с помощью сохраненных скейлеров. Затем модель загружается из файла и применяется для прогнозирования модуля упругости при растяжении на основе введенных пользователем данных.

Пользователю предоставляется возможность выбрать между выполнением прогноза и выходом из приложения. Если пользователь выбирает прогноз, то результаты прогнозирования выводятся на экран. Если пользователь выбирает выход, то приложение завершает свою работу. Если при вводе данных происходит ошибка, пользователю предлагается повторить операцию.

Таким образом, данное приложение является полезным инструментом для специалистов материаловедения и механики, и позволяет быстро и удобно получать прогнозы модуля упругости при растяжении. 

Рисунок 22 – Приложение прогнозирования модуля упругости при растяжении

* 1. **Создание удаленного репозитория**

Создание открытого репозитория на веб-сервисе GitHub позволит другим специалистам в области Data Science получить доступ к результатам и использовать их для своих проектов.

В репозиторий были загружены следующие файлы: Jupyter Notebook с кодом и результатами работы, презентация, пояснительная записка и файл README с кратким описанием проекта и инструкцией по его запуску. Адрес репозитория: <https://github.com/elenasamsonenko/predicting_composite>

Репозиторий был сделан открытым, что позволяет другим пользователям просматривать и скачивать его содержимое без ограничений (рисунок 23).

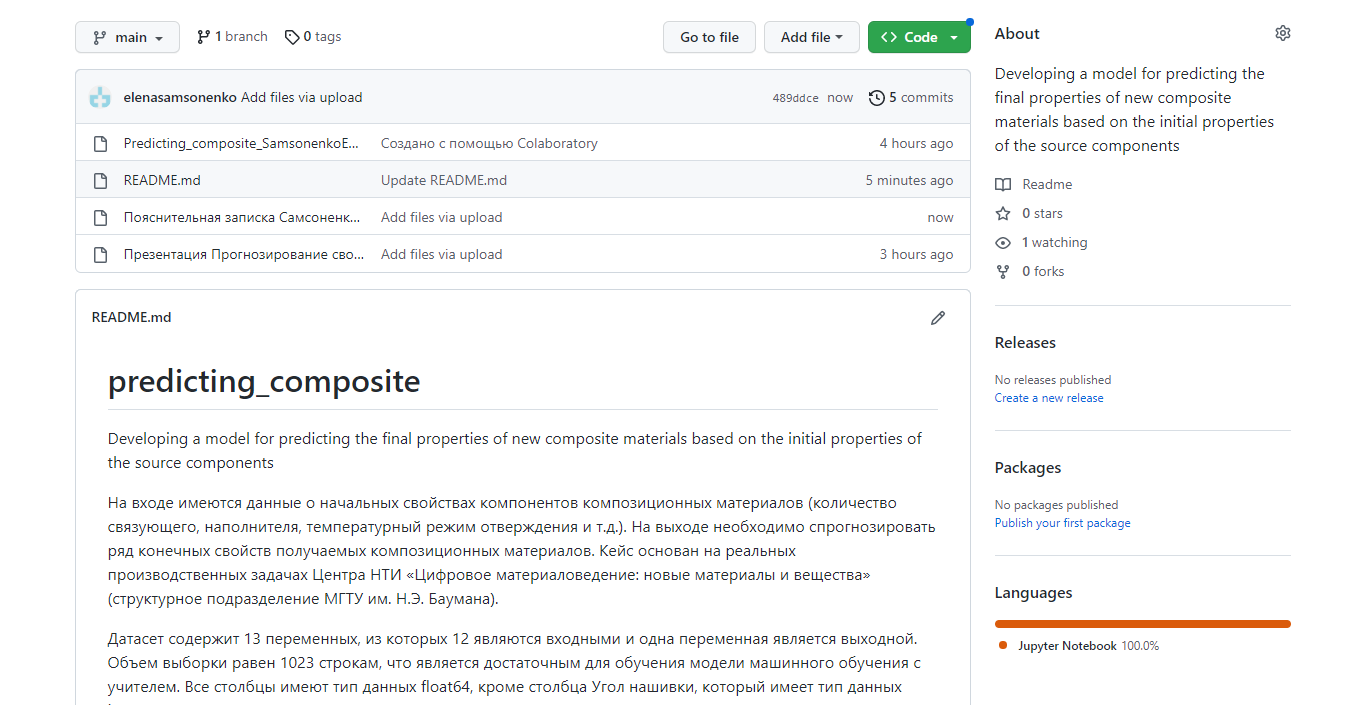


Рисунок 23 – Репозиторий на GitHub

Открытые репозитории позволяют делиться кодом и материалами с другими людьми и сообществами. Это помогает получить обратную связь и поддержку от коллег, улучшить навыки и знания. Также можно демонстрировать свои работы потенциальным работодателям или клиентам и наладить контакты с другими разработчиками.

['Модуль упругости при растяжении, ГПа'] вызов модели [[0.49884585]]

# **Заключение**

В заключении аналитической работы можно отметить, что задача прогнозирования конечных свойств композиционных материалов является сложной и актуальной и требует более тщательного и глубокого исследования.

В ходе выполнения работы были проведены различные этапы работы с данными, такие как анализ данных, предобработка и нормализация данных, построение нескольких моделей машинного обучения, таких как линейная регрессия, Лассо, метод опорных векторов, деревья решений, случайный лес, стохастический градиентный спуск и модель к-ближайших соседей. А также были обучены нейронные сети: многослойный персептрон и Sequential.

Однако все модели, включая нейронные сети, не показали хороших результатов по метрикам качества, что говорит о том, что для достижения более точных прогнозов необходимо обратиться за помощью к экспертам в области композитных материалов для получения дополнительных знаний о свойствах и особенностях материалов, а также для выбора и настройки моделей машинного обучения. Возможно также увеличить объем доступных данных, в том числе провести дополнительные физические испытания для уточнения значений переменных и увеличения корреляции между ними.

В целом выполнение данной аналитической работы позволило получить ценный опыт в работе с данными, построении моделей машинного обучения и нейронных сетей, оценке их качества, а также создании приложений с интерфейсом командной строки и размещении кода в репозитории. Это позволит другим специалистам в области композитных материалов использовать результаты работы для создания новых материалов с оптимальными свойствами в будущем. Таким образом, данная работа вносит свой вклад в развитие науки о материалах и промышленности, а также демонстрирует возможности и перспективы применения машинного обучения и нейронных сетей для решения сложных задач прогнозирования свойств материалов.

**Библиографический список**

1. Мюллер, А., Гвидо, С. Введение в машинное обучение с помощью Python [Текст] : руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо ; пер. с англ. А. А. Калинина. – М. : ДМК Пресс, 2017. – 416 с.
2. Бринк, Х., Ричардс, Дж., Феверолф, М. Машинное обучение [Текст] : учебное пособие / Х. Бринк, Дж. Ричардс, М. Феверолф ; под общ. ред. Н. И. Тихонова. – СПб. : Питер, 2017. – 352 с.
3. Лопес де Прадо, М. Машинное обучение: алгоритмы для бизнеса [Текст] / М. Лопес де Прадо ; пер. с англ. А. В. Кузнецова и др. – М. : Альпина Паблишер, 2019. – 400 с.
4. Шолле, Ф. Глубокое обучение на Python [Текст] : основы теории и практики / Франсуа Шолле ; перевод с английского Евгения Ткаченко и др… – СПб.: Питер, 2018 – 416 с.
5. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение [Текст] : учебник / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; перевод с английского А. В. Логунова и др… – Москва : ДМК Пресс, 2018. – 656 с.
6. Шевляков, Г. Л. Линейные модели и линейная регрессия [Текст] : учебное пособие / Г. Л. Шевляков. – М. : Изд-во МГУ, 2006. – 288 с.
7. Браунли, Дж. Подготовка данных для машинного обучения [Текст] : учебное пособие / Дж. Браунли ; перевод с англ. Е. С. Капацца ; под ред. А. В. Лебедева . – Москва : Издательство МГУ, 2021 . – 250 с.
8. Жерон Орельен − Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow [Текст] : учебник / Орельен Жерон ; пер. с англ. А. Кузнецова и др. – Москва : ДМК Пресс, 2018. – 576 с.
9. Поздняков, С. Статистика, R и анализ данных [Текст] : учебное пособие / С. Поздняков ; под ред. Е. С. Капацца . – Москва : Издательство МГУ, 2021 . – 300 с.
10. Композиционные материалы [Текст] : учебник для вузов / под ред. А. А. Берлина. – Москва : Высшая школа, 2007. – 543 с.
11. Марк Лутц. Изучаем Python [Текст] : учебник для начинающих и профессионалов / Марк Лутц. – Москва : Вильямс, 2019. – 1504 с.
12. Эрик Мэтиз. Изучаем Python. Программирование игр, визуализация данных, веб-приложения [Текст] : учебник для начинающих / Эрик Мэтиз. – Москва : Питер, 2019. – 1280 с.
13. П. Брюс, Э. Брюс. Разведочный анализ данных [Текст] : учебник / П. Брюс, Э. Брюс. – СПб.: БХВ-Петербург, 2018. – 304 с.
14. Михнин А. А. Оценка качества моделей машинного обучения: выбор, интерпретация и применение метрик [Текст] : учебное пособие / А. А. Михнин. – Москва : Литмаркет, 2023. – 67 с.
15. Описательная статистика [Электронный ресурс] // Википедия : сайт . – Режим доступа: https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D0%B8%D1%81%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F\_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0 (дата обращения: 15.03.2023).
16. Нормальное распределение [Электронный ресурс] // Википедия : сайт . – Режим доступа: https://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5\_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5 (дата обращения: 23.03.2023).
17. DataStart. Seaborn Heatmaps: 13 способов настроить визуализацию матрицы корреляции [Электронный ресурс] // DataStart : сайт. – Режим доступа: https://datastart.ru/blog/read/seaborn-heatmaps-13-sposobov-nastroit-vizualizaciyu-matricy-korrelyacii. (дата обращения: 22.03.2023).
18. CodeCamp. Как читать корреляционную матрицу [Электронный ресурс] // CodeCamp : сайт. – Режим доступа: https://www.codecamp.ru/blog/how-to-read-a-correlation-matrix/. (дата обращения: 22.03.2023).
19. Z-оценка [Электронный ресурс] // Википедия : сайт. – Режим доступа: https://ru.wikipedia.org/wiki/Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0 (дата обращения: 21.03.2023).
20. Как вычислить Z оценку: 15 шагов (с иллюстрациями) [Электронный ресурс] // WikiHow : сайт . – Режим доступа: https://ru.wikihow.com/%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D1%8C-Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D1%83 (дата обращения: 20.03.2023).
21. sklearn.preprocessing.MinMaxScaler [Электронный ресурс] // scikit-learn 1.2.2 documentation : сайт . – Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (дата обращения: 19.03.2023).
22. Нормализация данных в Python [Электронный ресурс] // Pythonist : сайт . – Режим доступа: https://pythonist.ru/normalizacziya-dannyh-v-python/ (дата обращения: 19.03.2023).
23. seaborn.pairplot [Электронный ресурс] // seaborn 0.12.2 documentation : сайт . – Режим доступа: https://seaborn.pydata.org/generated/seaborn.pairplot.html (дата обращения: 18.03.2023).
24. Creating Pair Plots in Seaborn with sns pairplot [Электронный ресурс] // datagy : сайт . – Режим доступа: https://datagy.io/seaborn-pairplot/ (дата обращения: 18.03.2023).
25. График плотности [Электронный ресурс] // DataVizCatalogue : сайт . – Режим доступа: https://datavizcatalogue.com/RU/metody/grafik\_plotnosti.html (дата обращения: 21.03.2023).
26. Плотность вероятности [Электронный ресурс] // Википедия : сайт . – Режим доступа: https://ru.wikipedia.org/wiki/%D0%9F%D0%BB%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C\_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B8 (дата обращения: 18.03.2023).
27. Scraper. Случайные леса и деревья решений с нуля в питоне [Электронный ресурс] : – Режим доступа: https://machinelearningmastery.ru/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249/. (дата обращения: 22.03.2023).
28. Википедия. Метод k-ближайших соседей [Электронный ресурс] : – Режим доступа: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4\_k-%D0%B1%D0%BB%D0%B8%D0%B6%D0%B0%D0%B9%D1%88%D0%B8%D1%85\_%D1%81%D0%BE%D1%81%D0%B5%D0%B4%D0%B5%D0%B9. (дата обращения: 22.03.2023).
29. Стохастический градиентный спуск [Электронный ресурс] : статья / Википедия. – Режим доступа: https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%BE%D1%85%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9\_%D0%B3%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82%D0%BD%D1%8B%D0%B9\_%D1%81%D0%BF%D1%83%D1%81%D0%BA. (дата обращения: 22.03.2023).
30. Git - Работа с удалёнными репозиториями [Электронный ресурс] : статья / Git-scm.com. – Режим доступа: https://git-scm.com/book/ru/v2/%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D1%8B-Git-%D0%A0%D0%B0%D0%B1%D0%BE%D1%82%D0%B0-%D1%81-%D1%83%D0%B4%D0%B0%D0%BB%D1%91%D0%BD%D0%BD%D1%8B%D0%BC%D0%B8-%D1%80%D0%B5%D0%BF%D0%BE%D0%B7%D0%B8%D1%82%D0%BE%D1%80%D0%B8%D1%8F%D0%BC%D0%B8(дата обращения: 26.03.2023)