МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Тема: «Прогнозирование конечных свойств новых материалов (композиционных материалов)»

Слушатель Харин Сергей Семенович

Москва, 2023

Содержание

1. Аналитическая часть
   1. Постановка задачи.
   2. Описание используемых методов
   3. Разведочный анализ данных
2. Практическая часть
   1. Предобработка данных
   2. Разработка и обучение модели
   3. Тестирование модели
   4. Написать нейронную сеть, которая будет рекомендовать соотношение матрица.
   5. Разработка приложения
   6. Создание удаленного репозитория и загрузка
   7. Заключение
   8. Библиографический список

**Введенение**

Композитный материал - это многокомпонентный материал, изготовленный (искусственно или естественным путем) из двух или более компонентов со значительно отличающимися физическими и/или химическими свойствами, которые объединяются для получения нового материала со свойствами, отличными от свойств отдельных компонентов, а не просто наложенными друг на друга. В композитных материалах принято различать матрицу/матрицу и наполнитель/наполнитель, последний выполняет армирующую функцию (по аналогии с армированием в композитных строительных материалах, таких как железобетон). Наполнителем в композитных материалах обычно являются углеродные или стеклянные волокна, а матрицей - полимер. Комбинация различных компонентов улучшает свойства материала, делая его одновременно легким и прочным. В то же время отдельные компоненты остаются нетронутыми в структуре композита, что отличает его от смесей и затвердевших строительных растворов. Изменяя состав, соотношение и ориентацию матрицы и наполнителя, можно получать различные материалы с требуемыми свойствами. Многие композитные материалы превосходят по своим механическим свойствам обычные материалы и сплавы. Использование композитных материалов обычно позволяет снизить вес конструкции при сохранении или улучшении ее механических свойств.

Разработка композитных материалов имеет очень высокий уровень сложности из-за огромного количества компонентов и условий обработки, так что изменение одного элемента может иметь множество непредвиденных последствий из-за их взаимодействия. Традиционные подходы к разработке новых материалов - это метод проб и ошибок, который приводит к проблемам низкой эффективности, высокой стоимости и неустойчивым результатам. С другой стороны, многочисленные эксперименты и вычислительные испытания накопили огромное количество многомерных, сложных и плохо изученных данных, которые могут скрывать важные правила структура-свойства. Машинное обучение может помочь выявить такие взаимосвязи и обеспечить возможность тестирования и оптимизации нескольких метрик одновременно, ускоряя открытие и оптимизацию дизайна материалов и прогнозируя композиции с желаемыми и уникальными свойствами.

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

Наша задача заключается в том, чтобы разработать модели для прогнозирования модуля упругости при растяжении, прочности при растяжении и соотношения матрица-наполнитель. У нас есть два файла с наборами данных.

Цель нашей работы - продемонстрировать наши знания и навыки в области Data Science, машинного обучения, баз данных, нейросетей и больших данных, которые мы получили в процессе обучения на квалификационном курсе. Для достижения этой цели мы должны выполнить следующие задачи:

- Провести разведочный анализ предложенных данных и предварительно обработать данные.

- Обучить несколько моделей для прогнозирования модуля упругости при растяжении и прочности при растяжении. Оценить точность моделей на тренировочных и тестовых наборах данных.

- Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Оценить точность прогноза на тренировочном и тестовом наборах данных.

- Разработать приложение с интерфейсом командной строки, которое будет предоставлять один или два прогноза.

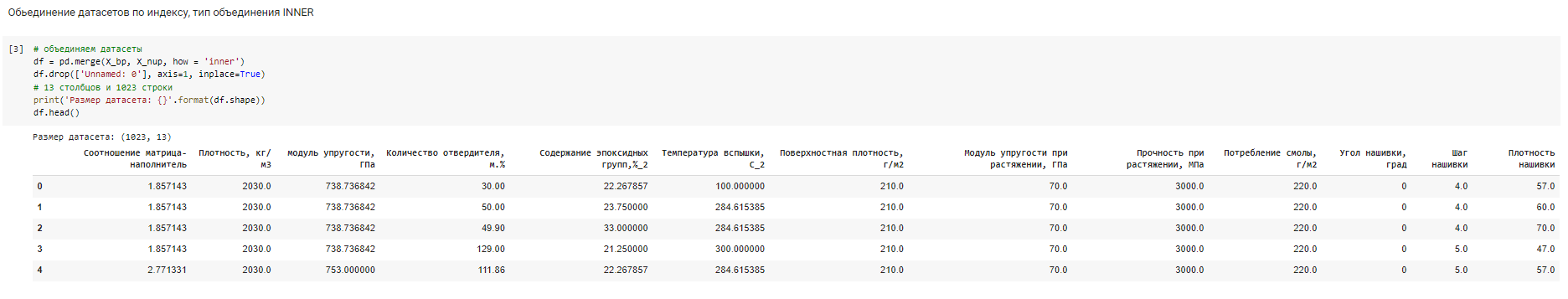
Для дальнейшей работы два файла с датасетами были объединены в один. Полученный датасет содержит 1023 строки и 13 столбцов (рисунок 1). Это означает, что часть данных (а именно 17 строк из датасета dataset\_nup) была удалена из таблицы и исключена из дальнейшего исследования.

Рисунок 1. Объединенный датасет для работы

* 1. **Описание используемых методов**

Всего в работе используется 7 моделей машинного обучения :

* Метод К-ближайших соседей
* Линейная регрессия
* Случайный лес
* Метод опорных векторов
* Градиентный бустинг
* Дерево решений
* Lasso регрессия.

**Метод К-ближайших соседей** заключается в том, что мы можем классифицировать образ, определив к какому классу он принадлежит на основе ближайшего к нему соседа. Для этого мы определяем k ближайших соседей для образа x и затем присваиваем ему класс, к которому относится наибольшее число образов из этой группы. Преимущества этого алгоритма заключаются в том, что он прост и понятен, легко обучается на новых данных, может работать с любым количеством категорий в задачах классификации, принимает два параметра - k и метрику расстояния (как правило, это эвклидово расстояние) и имеет низкую чувствительность к выбросам. Однако, его минусы состоят в том, что он имеет высокую стоимость вычислений, так как вам нужно обработать весь объем данных, и не работает так хорошо с категорическими параметрами.

**Линейная регрессия** является алгоритмом обучения с учителем в машинном обучении, который требует указания как входных, так и заранее подготовленных выходных данных для обучения модели. Эти данные вместе называются обучающей выборкой. Преимуществом линейной регрессии является ее простота в реализации. Однако, следует учитывать, что выбросы могут сильно повлиять на результаты регрессии.

**Случайный лес** - это алгоритм классификации, который состоит из многих деревьев решений (ансамбль решающих деревьев) и использует бэггинг и случайность признаков при построении каждого дерева, чтобы создать некоррелированный лес, прогноз которого точнее, чем у отдельного дерева. Преимущества этого метода в том, что он имеет высокую точность предсказания и практически не чувствителен к выбросам, масштабированию значений признаков и монотонным преобразованиям, не требует тщательной настройки параметров и хорошо работает «из коробки». Недостатки заключаются в том, что результаты случайного леса сложнее интерпретировать, алгоритм работает хуже многих линейных методов в случае большого количества разреженных признаков, склонен к переобучению на некоторых задачах, особенно на зашумленных данных, и предвзят в пользу признаков с большим количеством уровней, что может привести к неправильным предсказаниям на данных, содержащих категориальные переменные с различным количеством уровней.

**Метод опорных векторов** применяется для задач классификации и регрессии, а не только для регрессии, как было указано в вашем сообщении. SVM ищет оптимальную гиперплоскость в p-мерном пространстве, которая максимально разделяет два класса объектов, или предсказывает целевую переменную в регрессии. SVM работает путем поиска оптимальной разделяющей гиперплоскости, которая максимизирует зазор между двумя классами или уменьшает ошибку предсказания целевой переменной в регрессии. SVM имеет ряд гиперпараметров, которые могут быть настроены для достижения оптимальной производительности. Основные преимущества SVM включают в себя высокую точность предсказания, возможность работать с высокоразмерными данными, возможность настройки гиперпараметров для достижения оптимальной производительности и хорошую обобщающую способность. Однако, недостатки SVM заключаются в его чувствительности к выбросам и шуму в данных, а также в высокой вычислительной сложности для больших наборов данных.

**Градиентный бустинг** - это метод машинного обучения, основанный на последовательном построении нескольких базовых классификаторов, каждый из которых компенсирует ошибки предыдущего, чтобы получить лучший финальный классификатор. Он имеет ряд преимуществ, таких как учение на ошибках, выбор наблюдений на основе ошибки, простая настройка и легкая интерпретация. Однако, чтобы избежать переобучения, необходимо тщательно выбирать критерии остановки, и наблюдения с наибольшей ошибкой могут появляться чаще. Кроме того, он менее гибок, чем нейронные сети.

Метод **дерева решений** использует древовидную структуру для принятия решений на основе выбранных критериев. Алгоритм начинается с корневого узла, который представляет всю выборку, и на каждом шаге выбирает лучший признак для разделения выборки на подмножества. Процесс продолжается рекурсивно для каждого подмножества, пока не будет достигнут критерий остановки, например, когда все объекты в подмножестве относятся к одному классу или достигнуто максимальное число уровней дерева. Метод дерева решений имеет ряд преимуществ, таких как легкая интерпретируемость, возможность обработки данных различных типов и работа с отсутствующими данными. Кроме того, он может использоваться как для задач классификации, так и для задач регрессии. Однако метод дерева решений также имеет некоторые недостатки. Например, при построении больших деревьев может возникнуть переобучение, когда модель хорошо подстраивается под обучающую выборку, но плохо обобщается на новые данные. Кроме того, он может быть неустойчив к малым изменениям в данных, что может привести к значительным изменениям в структуре дерева.

Метод **Lasso регрессия** - это метод регуляризации, используемый для выбора наиболее значимых признаков в модели линейной регрессии. Он основан на добавлении штрафа L1-нормы к функции потерь модели, что приводит к уменьшению коэффициентов признаков до нуля, тем самым выбирая только наиболее значимые признаки. Для обучения модели Lasso регрессии используется метод наименьших квадратов, при этом функция потерь модели выглядит как сумма квадратов разностей между предсказанными значениями и истинными значениями целевой переменной, умноженных на коэффициент регуляризации alpha, умноженный на L1-норму вектора весов модели. Таким образом, функция потерь Lasso регрессии выглядит следующим образом: Loss = Sum of squared differences between predicted and true values + alpha \* L1-norm of weights. Преимущества метода Lasso регрессии включают возможность выбора наиболее значимых признаков в модели, что может привести к более простой и интерпретируемой модели. Кроме того, Lasso регрессия может использоваться для снижения эффекта мультиколлинеарности между признаками, что может улучшить качество модели. Однако, метод Lasso регрессии также имеет некоторые недостатки. Например, если признаки взаимосвязаны, то Lasso регрессия может выбрать только один из них, что может привести к потере информации и снижению качества модели. Кроме того, подбор оптимального значения параметра alpha может быть сложным и требовать значительного времени и ресурсов.

* 1. **Разведочный анализ данных**

Разведочный анализ данных является необходимой процедурой для получения представления о распределении данных и оценки качества данных, включая пропуски и выбросы. В ходе разведочного анализа учитываются и сравниваются множество признаков и закономерностей. Без понимания характера данных и их качества трудно работать с ними, поэтому важно изучить набор данных перед удалением или изменением каких-либо значений. Загрузка данных и вывод некоторых статистик позволяют получить необходимое представление о данных.

Une image contenant table

Description générée automatiquement

Рисунок 2. Информация о столбцах и значениях датасета

Une image contenant texte

Description générée automatiquementДействительно, в разведочном анализе применяется множество методов для оценки характеристик датасета, включая гистограммы распределения, диаграммы ящика с усами, попарные графики рассеяния точек, тепловую карту, а также анализ и удаление выбросов, пропусков и дубликатов. Эти методы помогают понять, как данные распределены, выявить потенциальные проблемы в данных, такие как отсутствие значений, необычные выбросы или ошибки в данных. После проведения разведочного анализа можно сделать более информированные выводы о данных и выбрать соответствующие методы анализа и моделирования для достижения нужных целей.

Рисунок 3. Поиск пропусков в датасете

Une image contenant texte

Description générée automatiquement

Рисунок 4. Поиск дупликатов в датасете

Une image contenant texte

Description générée automatiquement

Рисунок 5. Поиск уникальных значений в датасете

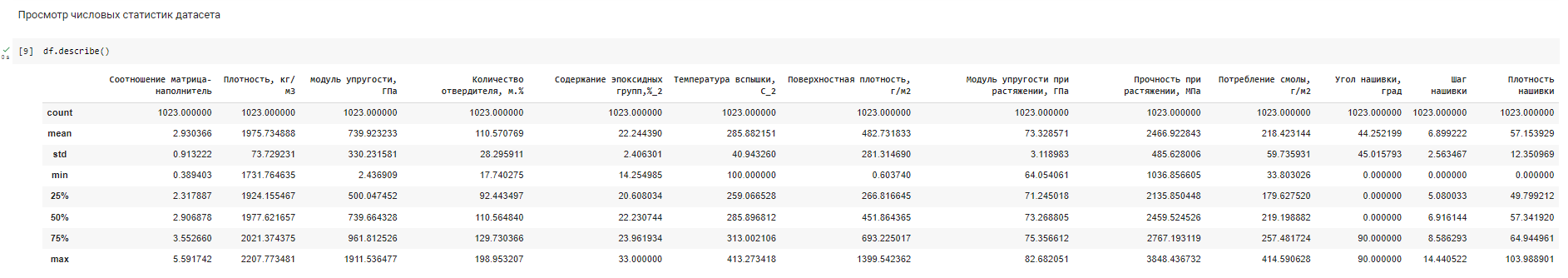


Рисунок 6. Описательная статистика датасета

Une image contenant table

Description générée automatiquement

Рисунок 7. Описательная статистика датасета (2)

Une image contenant graphique

Description générée automatiquement

Рисунок 8. Гистограммы и ящики с усами

Гистограмма является удобным инструментом визуализации, который используется для представления распределения одной или нескольких переменных. Она строится путем разбиения значений на интервалы и подсчета количества наблюдений, попавших в каждый интервал. Гистограмма позволяет оценить различные характеристики распределения, такие как центральную тенденцию, дисперсию, форму и симметрию. Она также может использоваться для оценки нормальности эмпирического распределения. Кроме того, на гистограмму можно накладывать кривую распределения, чтобы визуально сравнить эмпирическое распределение с теоретическим распределением. Это помогает оценить, насколько хорошо данные соответствуют определенному теоретическому распределению и использовать эту информацию для выбора соответствующей модели.

Исследование графиков показало, что большинство параметров в датасете имеют нормальное распределение, за исключением «Угла нашивки» и «Поверхностной плотности, г/м2». Поскольку значения «Угла нашивки» ограничены двумя значениями (0 и 1), он считается бинарным и категориальным признаком.

"Ящик с усами" - это диаграмма, используемая в статистическом анализе для компактного представления одномерного распределения вероятностей. Она показывает медиану, нижний и верхний квартили, минимальное и максимальное значения и выбросы.

Выбросы - это точки данных, значительно отличающиеся от общей выборки и могут вызывать проблемы при статистическом анализе. На диаграмме "ящик с усами" выбросы выделяются как точки, находящиеся за границами нижнего и верхнего усов. В данном случае, выбросы были обнаружены во всех параметрах, кроме "Угла нашивки".

Выбросы могут исказить данные и повлиять на статистические характеристики модели, такие как среднее значение и дисперсия. Они также могут привести к потере точности модели. Поэтому необходимо обнаружить и удалить выбросы перед моделированием данных.

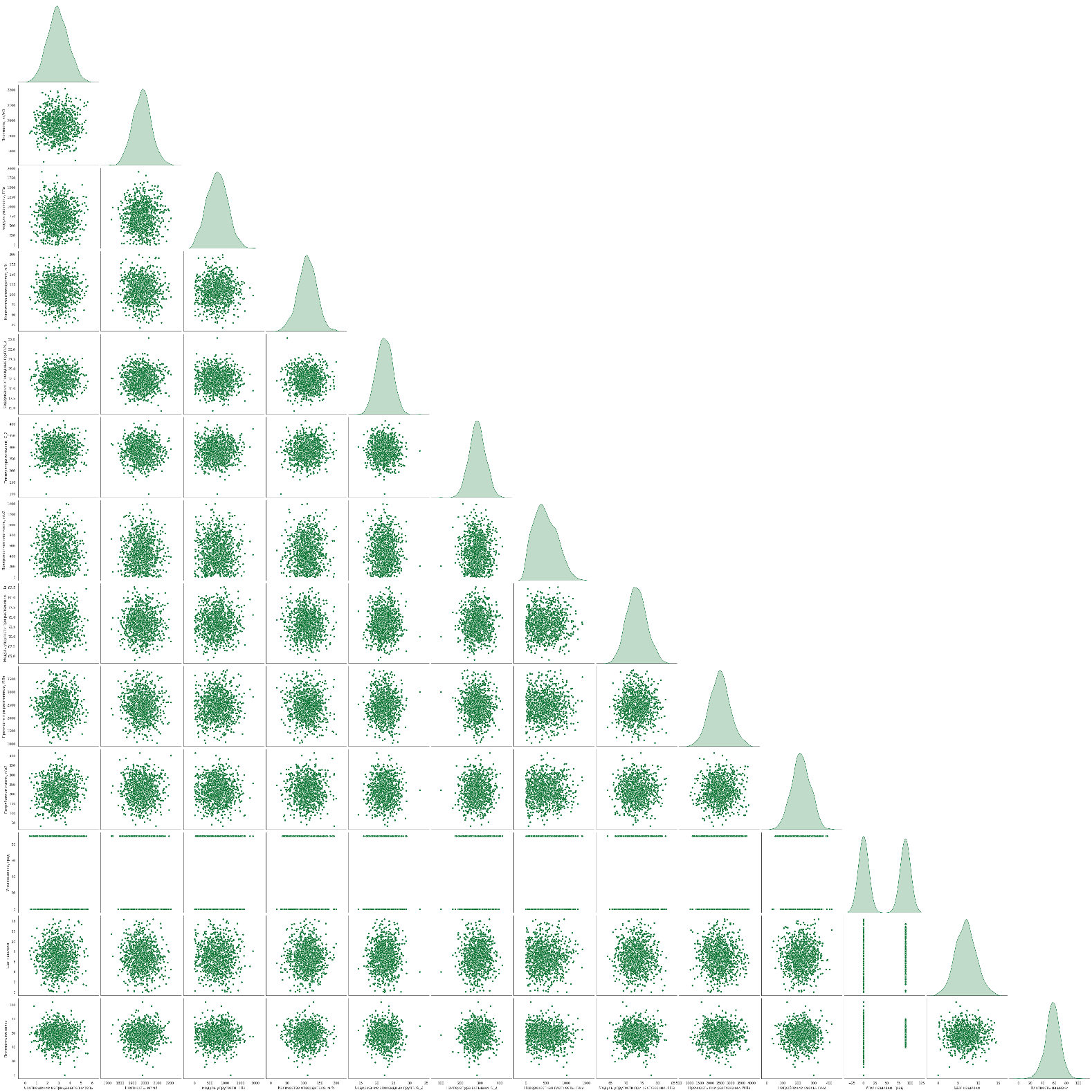


Рисунок 9. Гистограммы и ящики с усами

Согласно представленным графикам, не обнаружено линейной зависимости между характеристиками композитных материалов. Однако, необходимо провести анализ корреляций между переменными. Корреляционная связь подразумевает согласованные изменения между двумя или более признаками, где изменение одной переменной может вызвать закономерное изменение другой(-их) переменной(-ых).

Une image contenant graphique

Description générée automatiquement

Рисунок 10. Тепловая карта

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

Первым шагом предварительной обработкой рассмотрим удаление всех выбросов из данных до моделирования.

Une image contenant texte

Description générée automatiquement

Рисунок 11. Удаление выбросов

Une image contenant diagramme

Description générée automatiquement

Рисунок 12. Чистый датасет после трёх чисток

Тест Шапиро-Уилка позволяет определить, является ли распределение выборки нормальным. Если значение p-уровня значимости (обычно 0,05) ниже указанного порога, то можно утверждать, что данные выборки не имеют нормальное распределение. В данном случае, значение p-уровня значимости оказалось меньше 0,05 для следующих характеристик: модуль упругости в ГПа, поверхностная плотность в г/м2, модуль упругости при растяжении в ГПа, потребление смолы в г/м2 и угол нашивки. Это говорит о том, что распределение для данных характеристик не является нормальным.

Une image contenant texte

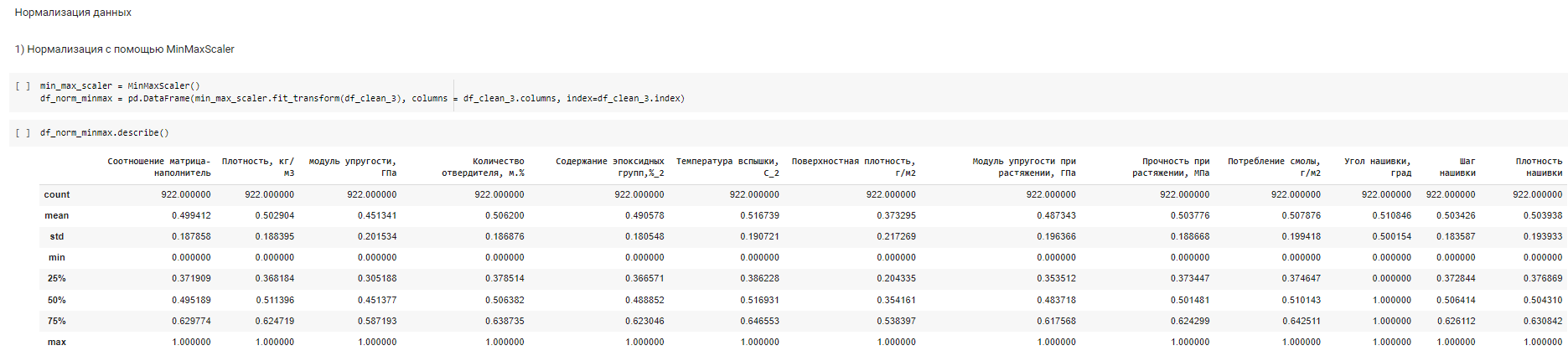
Description générée automatiquement

Рисунок 13. Тест Шапиро-Уилка

**Нормализация** - это процедура предварительной обработки входных данных (например, обучающих, тестовых, валидационных выборок или реальных данных), в которой значения признаков во входном векторе приводятся к определенному заданному диапазону, такому как [0...1] или [-1...1]. Главная цель нормализации - приведение данных, измеренных в разных единицах и диапазонах, к единому виду, который позволяет сравнивать данные и использовать их для расчета схожести объектов в выборке. Перед обучением модели все признаки должны быть приведены к одинаковому влиянию друг на друга. В библиотеке Scikit-learn на Python для этого доступны классы MinMaxScaler и RobustScaler.

**Стандартизация** представляет собой процесс приведения данных к определенному формату и представлению, что обеспечивает возможность их правильного использования в многомерном анализе. Она позволяет исключить влияние отклонений по каждому из признаков и привести все значения в датасете к нормальному распределению с математическим ожиданием, равным нулю, и стандартным отклонением, равным единице. Это приводит к созданию стандартизированной шкалы, которая определяет положение каждого значения в наборе данных и измеряет его отклонение от среднего в единицах стандартного отклонения.

Функция MinMaxScaler была использована для форматирования данных, что позволило привести все параметры к одинаковому относительному масштабу. При этом сохранены относительные различия между значениями каждого объекта. Значения каждого признака были масштабированы таким образом, что максимальное значение признака стало равным 1, а минимальное - 0.



Une image contenant graphique

Description générée automatiquementРисунок 14. Описание датасета после преобразования с MinMaxScaler

Рисунок 15. Визуализация после преобразования с MinMaxScaler

* 1. **Разработка и обучение моделей**

Для создания модели машинного обучения, которая может прогнозировать характеристики композитных материалов, такие как модуль упругости при растяжении и прочность при растяжении, необходимо выполнить следующие этапы:

1. Нормализовать данные, чтобы гарантировать их однородность и сопоставимость.

2. Разделить нормализованные данные на обучающую и тестовую выборки, где 30% данных отводится для тестирования модели, а на оставшихся данных модель обучается.

Une image contenant texte

Description générée automatiquementДля моделей используем датасет, преобразованный с помощью MinMaxScaler.

Рисунок 16. Разбивка на тестовую и обучающую выборки

* Une image contenant texte, lettre

  Description générée automatiquementМетод К-ближайших соседей :

Une image contenant texte, lettre

Description générée automatiquement

Рисунок 17. K-NN Results

* Linear Regression :

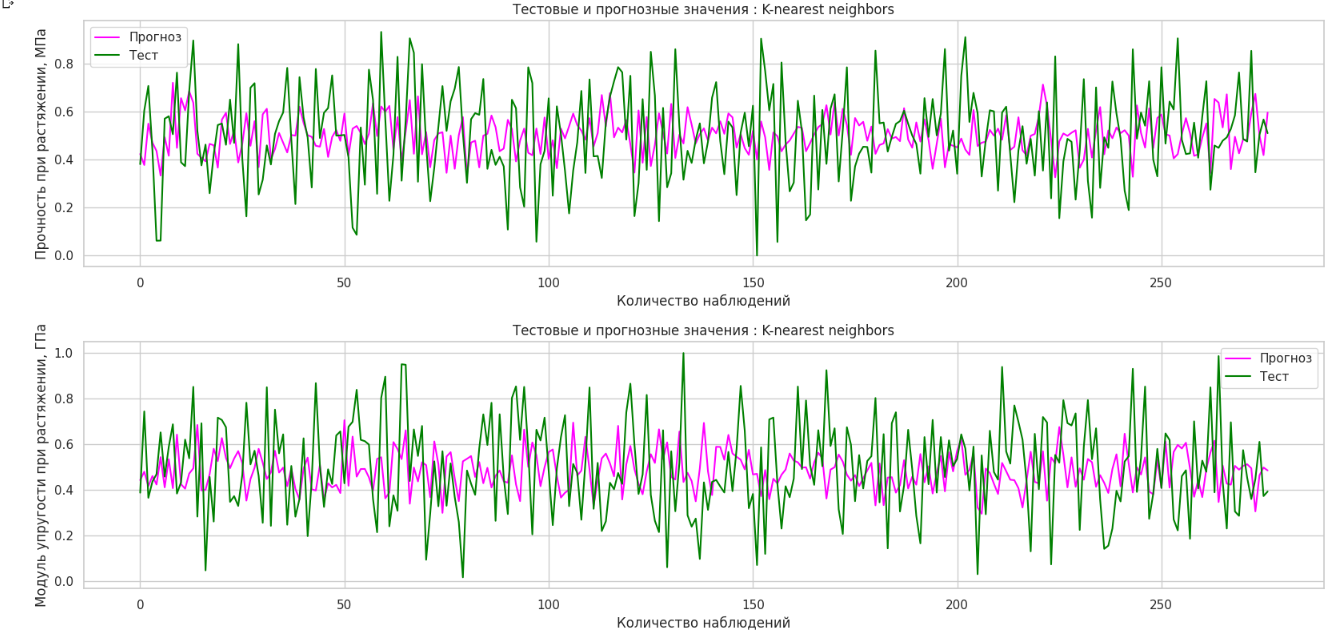


Рисунок 17. K-NN results

* Linear Regression

Une image contenant texte, lettre

Description générée automatiquement

Une image contenant texte, lettre

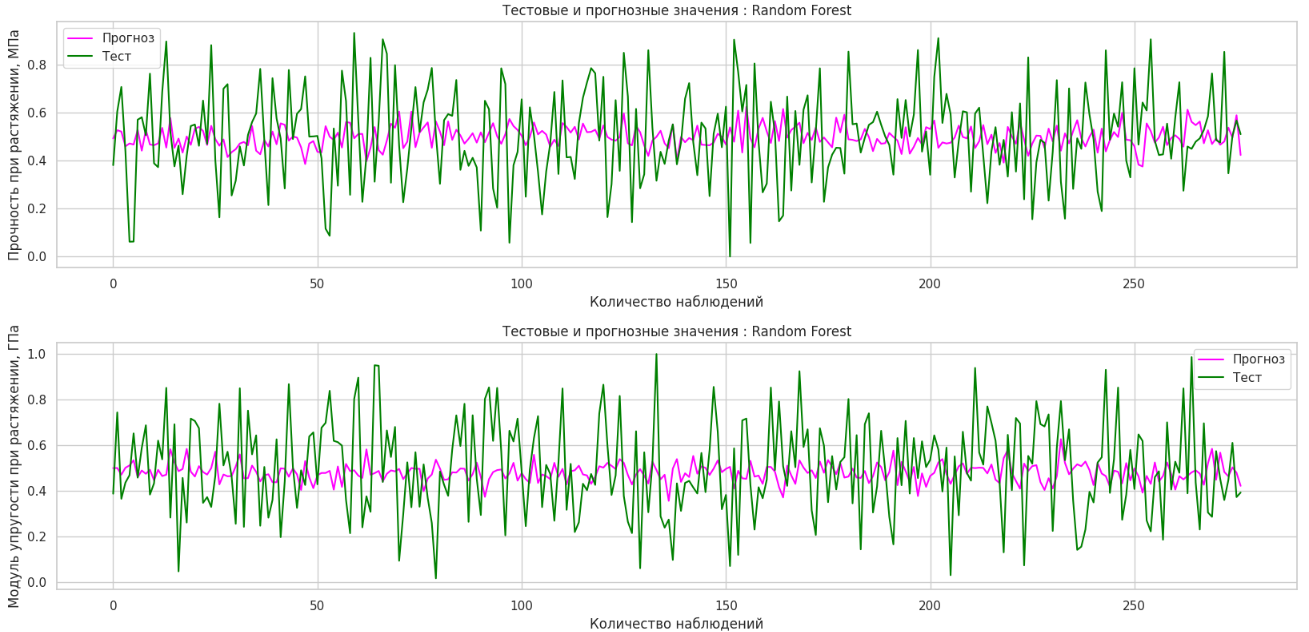
Description générée automatiquement



Рисунок 18. Linear Regression results

* Une image contenant texte

  Description générée automatiquementRandom Forest

Une image contenant texte, lettre

Description générée automatiquement

Рисунок 19. Random Forest results

* Support Vector

Une image contenant texte

Description générée automatiquement

Une image contenant texte, lettre

Description générée automatiquement

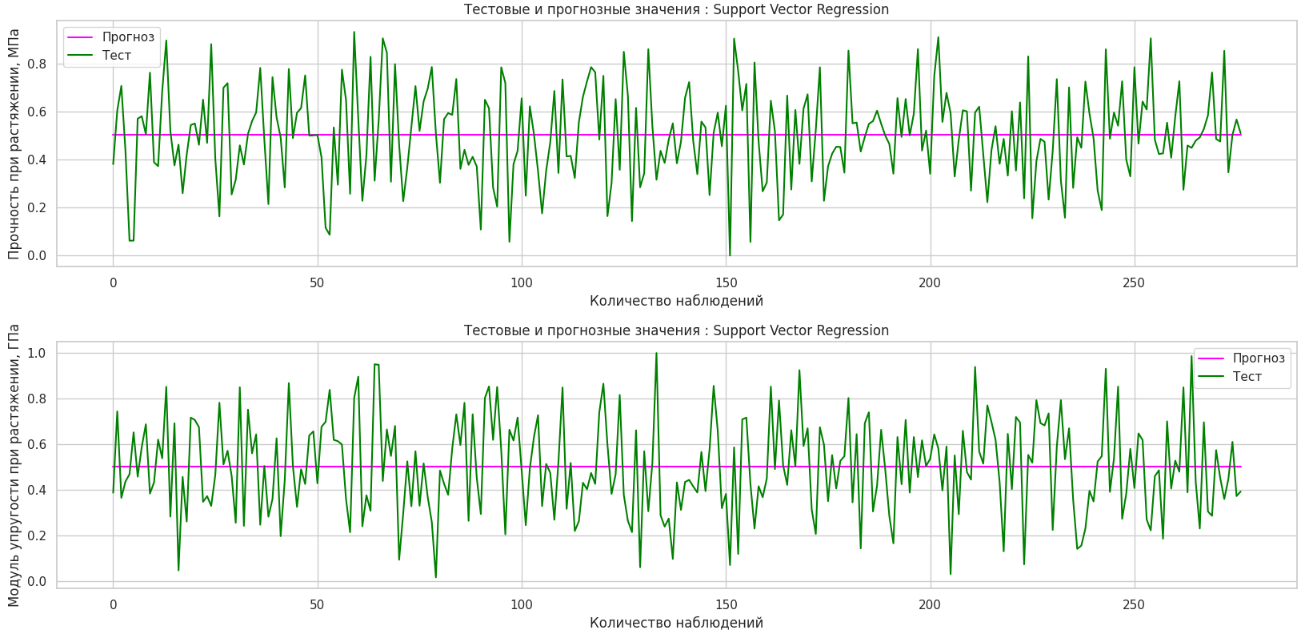


Рисунок 20. Support Vector results

* Gradient Boosting

Une image contenant texte, lettre

Description générée automatiquement

Une image contenant texte

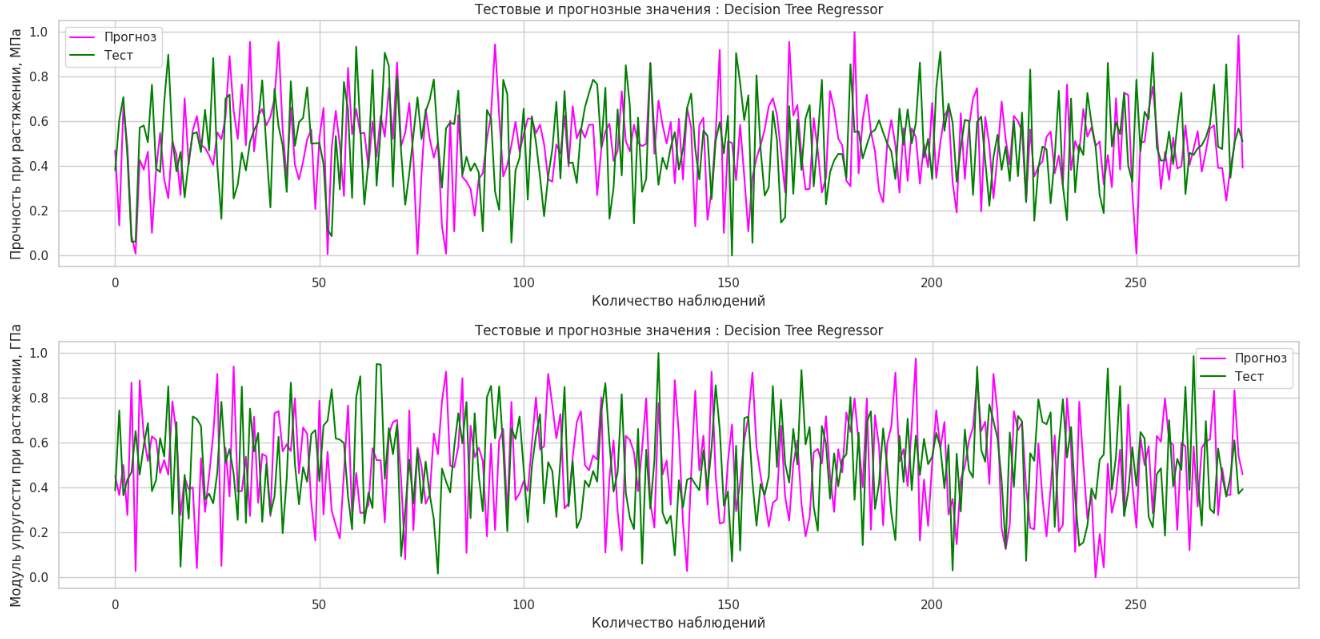
Description générée automatiquement

Рисунок 21. Gradient Boosting results

* Decision Tree Regressor

Une image contenant texte, lettre

Description générée automatiquement

Une image contenant texte, lettre

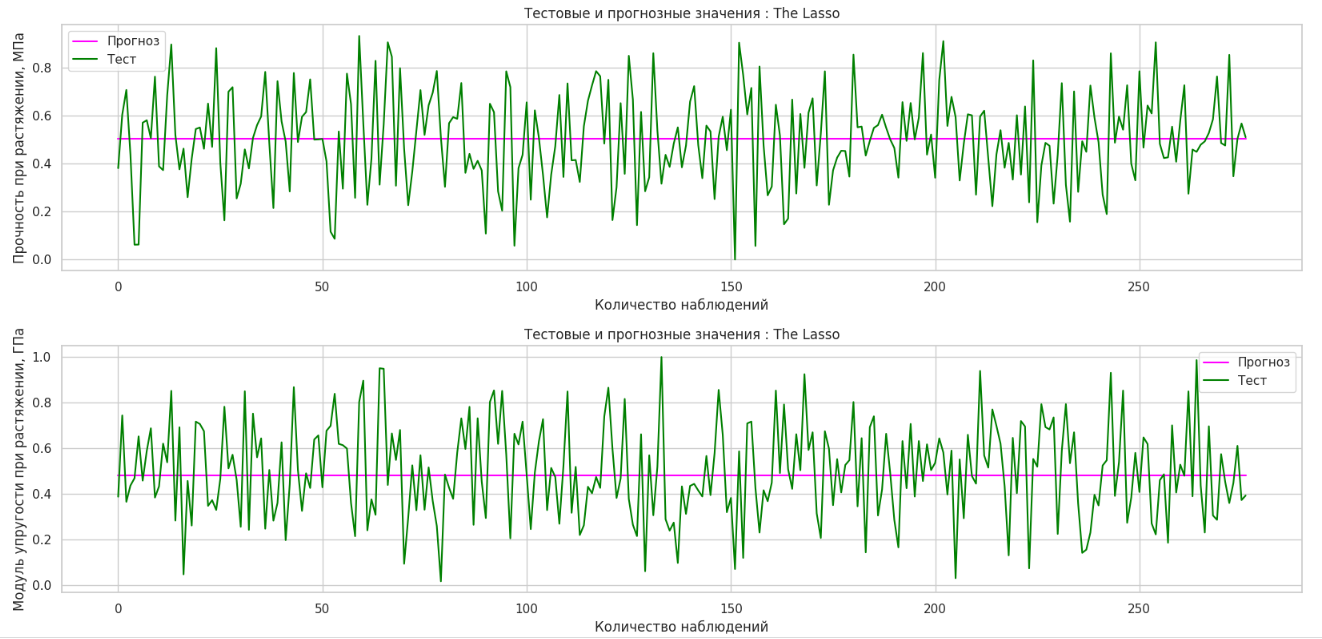
Description générée automatiquement

Рисунок 22. Decision Tree Regressor

* The Lasso

Une image contenant texte, lettre

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Рисунок 23. Lasso Regressor

Une image contenant table

Description générée automatiquementПриведём полную таблицу с результатами всех моделей :

Если коэффициент детерминации (R2) отрицателен, то это свидетельствует о низкой способности модели к обобщению. Такая модель работает хуже, чем простое вычисление среднего. Наилучшие результаты R2 и MAE на тестовой выборке были получены с помощью алгоритма регрессии «SVR».

Все модели, используемые для прогнозирования модуля упругости при растяжении, показали неудовлетворительный результат. Аналогично, при разработке моделей для прогнозирования прочности при растяжении были получены модели с низкой обобщающей способностью, которые также не решают поставленную задачу.

Попробуем прогнать датасет через модели ещё раз, но теперь с данными нормализованными с помощью Normalizer.

Normalizer в машинном обучении - это метод нормализации данных, который преобразует каждую строку данных (т.е. объект) в единичную норму (норму длины равной 1). Это означает, что каждый вектор признаков будет масштабирован таким образом, что его длина (норма) станет равной 1.

В простейшем случае, Normalizer может быть использован для нормализации каждой строки данных путем деления каждого значения признака на евклидову длину всего вектора признаков. Однако, Normalizer также может быть настроен для использования других норм, таких как Манхэттенская норма или норма Чебышева.

Une image contenant texte, blanc, capture d’écran

Description générée automatiquementПреимущества использования Normalizer включают устранение влияния различных масштабов и единиц измерения признаков в данных, что может повысить качество работы модели. Кроме того, Normalizer может уменьшить влияние выбросов в данных, так как он фокусируется на относительных значениях каждого признака, а не на абсолютных значениях.

Рисунок 24. Нормализация данных с помощью Normalizer

Une image contenant table

Description générée automatiquement

Рисунок 24. Результаты моделей на данных, преобразованных с помощью Normalizer

Модели дали на удивление очень хороший результат. Наилучший результат по R2 наблюдается у Gradient Boosting решрессора – 0.92 и 0.78.

Продолжим экперимент и прогоним датасет, преобразованный с помощью StandardScaler через модели.

Une image contenant table

Description générée automatiquement

Рисунок 25. Результаты моделей на данных, преобразованных с помощью StandardScaler

Результат снова неудовлетворительный, как и в случае с данными, нормализованными с помощью MinMaxScaler.

Une image contenant table

Description générée automatiquementПроделаем тот же опыт с непреобразованными данными, то есть, с изначальными данными без изменений и преобразований.

Рисунок 26. Результаты моделей на данных без преобразований

Результаты на данных без преобразований неудовлетворительны. Делаем вывод, что лучше всех модели работают на данных, нормализованных с помощью Normalizer.

* 1. **Нейронная сеть для рекомендации соотношения «матрица – наполнитель»**

Нейронные сети - это метод искусственного интеллекта, который обучает компьютеры обрабатывать данные так же, как и человеческий мозг, используя процесс машинного обучения, известный как глубокое обучение. Он состоит из взаимосвязанных узлов или нейронов, организованных в слоистую структуру, которая напоминает человеческий мозг. Это позволяет создавать адаптивные системы, которые учатся на своих ошибках и постоянно улучшают свою производительность. Таким образом, искусственные нейронные сети стремятся решать сложные задачи с большей точностью.

Обучение нейронной сети  - это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки.

Всего в работе было обучено более 20 моделей, на презентацию в работе было выбрано 5 следующих моделей.

Une image contenant texte

Description générée automatiquement

Рисунок 27. Предобработка данных на подачу в нейросеть

1. Модель 1

Une image contenant table

Description générée automatiquement

Рисунок 28. Создание слоёв и архитектуры модели

Une image contenant graphique

Description générée automatiquement

Рисунок 29. Визуализация ошибок модели



Рисунок 30. Результат работы модели

1. Модель 2

Une image contenant texte

Description générée automatiquement

Рисунок 31. Создание слоёв и архитектуры второй модели

Une image contenant graphique

Description générée automatiquement

Рисунок 32. Визуализация ошибок второй модели

Une image contenant graphique

Description générée automatiquement

Рисунок 33. Результат работы второй модели

1. Модель 3

Une image contenant table

Description générée automatiquement

Рисунок 34. Создание слоёв и архитектуры третьей модели

Une image contenant graphique

Description générée automatiquement

Рисунок 35. Визуализация ошибок третьей модели

Une image contenant graphique

Description générée automatiquement

Рисунок 36. Результат работы третьей модели

1. Модель 4

Une image contenant texte

Description générée automatiquement

Рисунок 37. Создание слоёв и архитектуры четвертой модели

Une image contenant graphique

Description générée automatiquement

Рисунок 38. Визуализация ошибок четвертой модели

Une image contenant graphique

Description générée automatiquement

Рисунок 38. Результат работы четвертой модели

1. Модель 5

Une image contenant texte

Description générée automatiquement

Рисунок 39. Создание слоёв и архитектуры пятой модели

Une image contenant graphique

Description générée automatiquement

Рисунок 40. Визуализация ошибок пятой модели

Une image contenant graphique

Description générée automatiquement

Рисунок 41. Результат работы пятой модели

Une image contenant table

Description générée automatiquement

Рисунок 42. Обьединенная таблица результатов всех моделей

Проанализировав полученные результаты, можно сделать следующие выводы. По метрике минимизации средней абсолютной ошибки (MAE) Model 3 и Model 5 справились лучше остальных, так как они имеют наименьшее значение MAE. По метрике минимизация среднеквадратической ошибки (MSE) Model 5 справилась лучше, так как имеет наименьшее значение MSE.

Также следует учитывать, что R2 (коэффициент детерминации) отражает соответствие модели данным, поэтому отрицательные значения этой метрики могут указывать на то, что модель не соответствует данным вообще, и ее не следует использовать.

Обобщая вывод, можер сказать, что Model 5 будет наилучшим выбором, так как имеет наименьшее значение MAE и R2 близок к нулю.

1. **Разработка приложения**

В работе было продела попытка создания web приложения и создание консольного приложения в Google Colab.

Une image contenant texte, lettre

Description générée automatiquementРассмотрим созданное консольное приложение.

Рисунок 43. Создание функции с переменными для приложения

Une image contenant texte

Description générée automatiquement

Рисунок 44. Создание функции приложения

Une image contenant texte, lettre

Description générée automatiquement

Рисунок 45. Результат работы консольного приложения

Une image contenant texte

Description générée automatiquementРассмотрим попытку создания web приложения на Flask.

Рисунок 46. Разработка приложения на Flask

Une image contenant texte, écran, capture d’écran, télévision

Description générée automatiquement

Рисунок 47. Разработка HTML страницы для приложения (1)

Une image contenant texte, écran, capture d’écran, plat

Description générée automatiquement

Рисунок 48. Разработка HTML страницы для приложения (2)

Une image contenant table

Description générée automatiquement

Рисунок 49. Запуск web приложения

1. **Заключение**

Работа с моделями прогнозирования, использующими машинное обучение, является сложным процессом, который требует как навыков программирования, так и профессиональных знаний в области композитных материалов. Чтобы обеспечить достоверность полученных прогнозов, необходимо иметь контакт с экспертами в данной области для консультаций. В ходе работы был проведен анализ датасета, построено множество графиков и выполнено разбиение данных на обучающую и тестовую выборки. Для реализации моделей машинного обучения были применены различные алгоритмы, включая метод К ближайших соседей, линейную регрессию, деревья решений, опорные вектора и случайный лес. Были составлены отчеты, оценивающие качество проводимого обучения, а также сравнение результатов работы моделей с помощью графиков и диаграмм. Была разработана нейронная сеть. Из результатов исследования стало понятно, что для повышения достоверности прогнозов необходимо дополнить датасет недостающими данными, а для этого необходимо иметь команду из специалистов по данной области для консультаций и сбора информации. Метрика, которую следует использовать для оценки прочности при растяжении, - метод градиентного бустинга, а для модуля упругости при растяжении - метод опорных векторов.

**Список литературы**

1. Apoorva, D. Регрессия (Regression) [Электронный ресурс] : – Режим доступа: https://www.helenkapatsa.ru/rieghriessiia/ (дата обращения: 14.04.2023).

2. Cheat code to find (MSE,RMSE,MAE)Mape [Электронный ресурс] : – Режим доступа: https://www.kaggle.com/code/udayreddie/cheat-code-to-find-mse-rmse-mae-mape (дата обращения: 11.04.2023).

3. Detect and Remove the Outliers using Python [Электронный ресурс] : – Режим доступа: https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/ (дата обращения: 07.04.2023).

4. Keras - последовательная модель Sequential [Электронный ресурс] : – Режим доступа: https://proproprogs.ru/tensorflow/keras-posledovatelnaya-model-sequential (дата обращения: 23.04.2023).

5. SciKeras 0.9.0 documentation. Migrating from tf.keras.wrappers.scikit\_learn [Электронный ресурс] : – Режим доступа: https://www.adriangb.com/scikeras/stable/migration.html (дата обращения: 24.04.2023).

6. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Обработка данных с использованием Python, SQL и машинного обучения [Текст] / Пер. с англ. – М.: ДМК Пресс, 2019. – 442 с.

7. Сравнительное изучение алгоритмов классического машинного обучения [Электронный ресурс] : – Режим доступа: https://machinelearningmastery.ru/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222/ (дата обращения: 13.04.2023).

8. Способы обнаружения и удаления выбросов [Электронный ресурс] : – Режим доступа: https://questu.ru/articles/434373/ (дата обращения: 07.04.2023).

9. Что такое нейронная сеть? [Электронный ресурс] : – Режим доступа: https://aws.amazon.com/ru/what-is/neural-network/ (дата обращения: 24.04.2023).

10. 2.3. Разведочный анализ данных (рад) [Электронный ресурс] : – Режим доступа: https://studfile.net/preview/8858767/page:8/ (дата обращения: 05.04.2023).

11. Документация Seaborn [Электронный ресурс]: – Режим доступа: https://seaborn.pydata.org/ (дата обращения: 21.04.2023).

12. Документация Flask [Электронный ресурс]: – Режим доступа: https://pypi.org/project/Flask/ (дата обращения: 21.04.2023).

13. Документация Plotly [Электронный ресурс]: – Режим доступа: https://pypi.org/project/Plotly/ (дата обращения: 21.04.2023).