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 Баркинхоева З.М.

Москва, 2022

**Содержание**

|  |  |
| --- | --- |
| Введение | 3 |
| Глава 1. Аналитическая часть | 5 |
| 1.1. Постановка задачи | 5 |
| 1.2. Описание используемых методов | 6 |
| 1.3. Разведочный анализ данных | 8 |
| Глава 2. Практическая часть | 10 |
| 2.1. Предобработка данных | 10 |
| 2.2. Разработка и обучение модели | 12 |
| 2.3. Тестирование модели | 12 |
| 2.4. Нейронная сеть для соотношения «матрица – наполнитель» | 13 |
| Заключение | 14 |
| Список используемых источников | 15 |

**Введение**

Композит – материал, изготовленный из двух или более компонентов с различными физическими или химическими свойствами. Это смесь, итоговые свойства которой превышают показатели каждого из ее элементов, взятого по отдельности: если сорвать с дерева лист и покрыть его смолой, получится композит: ни лист, ни смола сами по себе не будут так же держать форму или противостоять внешнему воздействию, как их сочетание – данный простой принцип и лег в основу материалов.[[1]](#footnote-1)

В 1796 году англичанин Д. Паркер запатентовал технологию получения романцемента – вяжущего материала, способного к затвердеванию и на открытом воздухе, и в воде. Смешанный с гравием, песком и водой цемент образовывал бетон. Бетон, отличавшийся своей пластичностью, завоевал новую популярность. Но прочный и водостойкий материал плохо растягивался, потому в качестве основного материала для несущих конструкций применялось железо в виде кованых стержней и полос. А оно было подвержено коррозии. Поэтому появилась задача найти более универсальный материал, способный сочетать свойства как бетона, так железа.

Так придумали железобетон: арматура легко обволакивалась бетоном, тем самым оказываясь включенной в его массу. Сила сцепления железа с бетоном была огромной: материалы начинали работать как одно целое. В бетоне железо не ржавело и сохраняло свою прочность, а сам железобетон показал высокую огнестойкость.[[2]](#footnote-2)

Сейчас видов композитов гораздо больше, но данный принцип сохраняется. Однако, у этого метода изготовления новых материалов есть и недостаток: если характеристики исходных компонентов и известны, это вовсе не значит, что определить характеристики конечного продукта будет просто. Для решения этой проблемы можно проводить физические испытания образцов материалов или прогнозировать характеристики.

Прогнозирование – симуляция представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Для текущего исследования, основная **цель** которого заключается в попытке спрогнозировать ряд конечных свойств получаемых композиционных материалов, на входе используются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.), предоставленные Центром НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана). В представленном кейсе речь идет о базальтопластике – полимерном композиционном материале, применение в котором базальтового наполнителя по сравнению со

стекловолокном позволяет снизить стоимость композиционного материала,

его гигроскопичность, повысить температуру эксплуатации.[[3]](#footnote-3)

**Актуальность** обусловлена необходимостью сократить количество проводимых физических испытаний.

**Глава 1. Аналитическая часть**

**1.1. Постановка задачи**

В ходе работы необходимо решить ряд последовательных **задач**:

1. Изучить теоретические основы и методы решения поставленной задачи.
2. Провести разведочный анализ предложенных данных.
3. Провести предобработку данных.
4. Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении.
5. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.
6. Оценить точность моделей на тренировочном и тестовом датасете.

Также нужно разработать приложение с графическим интерфейсом или интерфейсом командной строки с функцией прогноза. Полученные результаты размещаются в созданном репозиторий в GitHub / GitLab (https://github.com/zalinabarkin/VKR).

Для реализации поставленных задач даны 2 файла: X\_bp.xlsx (1024 строки и 11 столбцов) и X\_nup.xlsx (1040 строк и 4 столбцов). Удаление неинформативных столбцов и объединение таблиц по методу INNER привело к сокращению количества строк и столбцов в итоговом датасете до 1023 и 13, соответственно (то есть 17 строк из X\_nup.xlsx выпали).

**1.2. Описание используемых методов**

Поскольку данные в датасете относятся к непрерывным, в работу берутся регрессионные модели. И действительно впоследствии попытка обучения моделей Logistic Regression (которая требует категориальных признаков) и Decision Tree Classifier привели к ошибке «ValueError: Unknown label type: 'continuous'».

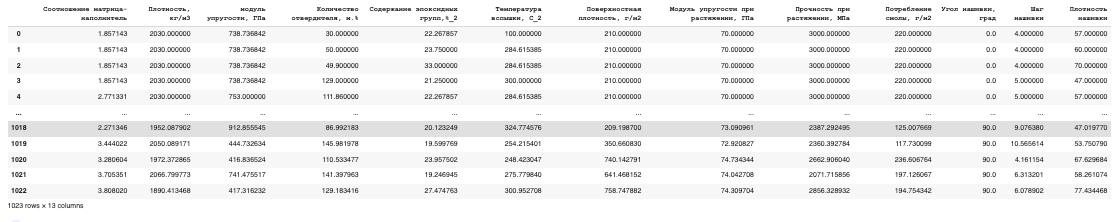
Регрессия – это технология статистического анализа, целью которой является определение лучшей модели, устанавливающей взаимосвязь между выходной (зависимой) переменной и набором входных (независимых) переменных.[[4]](#footnote-4)

В исследовании применены следующие методы:

1. Линейная регрессия – используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.
2. Регрессия ближайших соседей (Метод ближайших соседей, К ближайших соседей (kNN – k Nearest Neighbours) ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии). [[5]](#footnote-5)
3. Дерево решений (Decision Tree) – метод автоматического анализа больших массивов данных и эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Он помогает в решении задач по классификации и регрессии[[6]](#footnote-6). В данной работе используется дерево регрессии (Decision Tree Regressor), предназначенное для непрерывных целевых переменных, а не дерево классификации (Decision Tree Classifier), необходимое для дискретных переменных.

**1.3. Разведочный анализ данных**

Перед тем, как передать данные в модели, их необходимо изучить. Для этого проводится разведочный анализ – первичный осмотр данных с целью получения ценной информации: каких-либо зависимостей между атрибутами, зависимостей временного характера, ранжирование по степени влияния тех или иный атрибутов и т.д.[[7]](#footnote-7)



Проверка на пропуски демонстрирует их отсутствие (так же, как и повторений):

Соотношение матрица-наполнитель 0

Плотность, кг/м3 0

Модуль упругости, ГПа 0

Количество отвердителя, м.% 0

Содержание эпоксидных групп,%\_2 0

Температура вспышки, С\_2 0

Поверхностная плотность, г/м2 0

Модуль упругости при растяжении, ГПа 0

Прочность при растяжении, МПа 0

Потребление смолы, г/м2 0

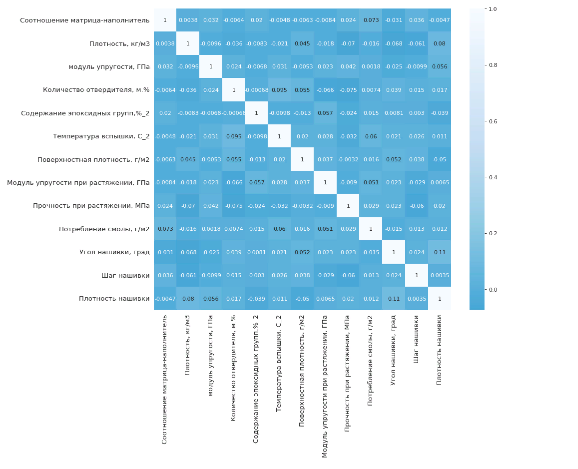
Угол нашивки, град 0

Шаг нашивки 0

Плотность нашивки 0

В качестве инструментов разведочного анализа используются оценка статистических характеристик датасета, гистограммы распределения каждой из переменной, диаграммы ящика с усами и попарные графики рассеяния точек.

Примененные методы анализа не выявили четкой зависимости, что и подтвердила тепловая карта с матрицей корреляции: результаты близки к 0, максимальный результат – 0,11.



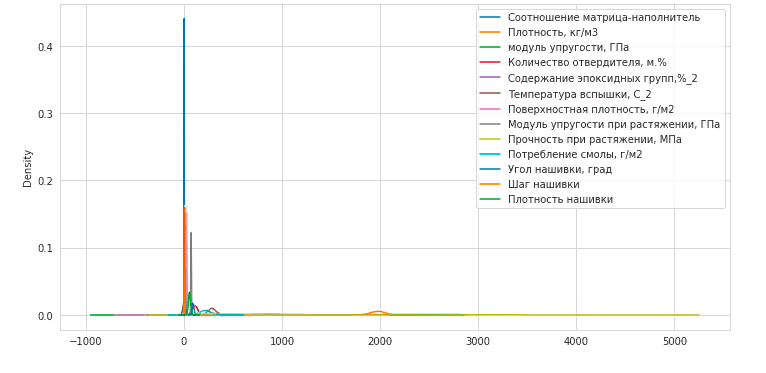
**Глава 2. Практическая часть**

**2.1. Предобработка данных**

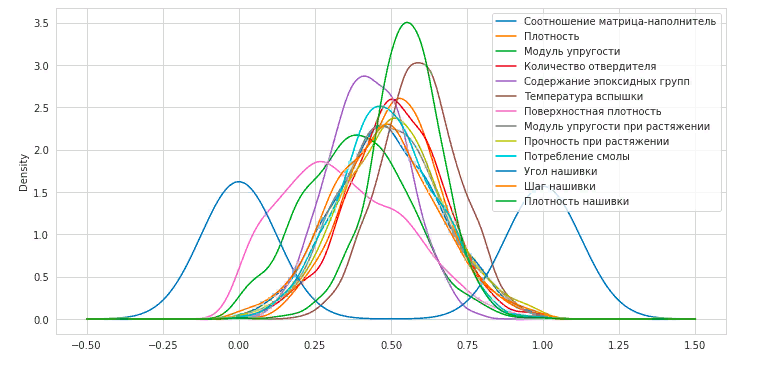
Условия поставленного задания требуют масштабирования данных путем нормализации. Нами выбран метод MinMaxScaler.



Визуализация данных до нормализации:



Визуализация данных после нормализации:



**2.2. Разработка и обучение модели**

Разработку и обучение моделей машинного обучения было решено провести отдельно для каждого из выходных параметров (по условию «Прочность при растяжении» и «Модуль упругости при растяжении»).

Также по условию задания при построении модели необходимо 30% данных оставлялось на тестирование модели, на остальных проводилось обучение моделей. Поэтому в коде test\_size равен 0.3.

**2.3. Тестирование модели**

После обучения моделей проведена оценка точности этих моделей на

обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя квадратическая ошибка (MSE) и коэффициент детерминации (R2 или R-квадрат), который показывает, какая доля изменчивости целевой переменной объясняется с помощью использованной модели.

*Прочность при растяжении*

|  |  |  |
| --- | --- | --- |
| Модель | MSE | R2 |
| Linear Regression | 0.028208863778752036 | 0.004230422816883017 |
| KNeighborsRegressor | 0.028208863778752036 | 0.004230422816883017 |
| Decision Tree Regressor | 0.028470633138137295 | -0.005010004814622437 |

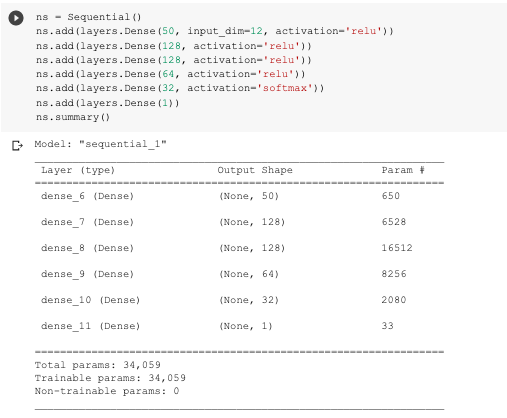
*Модуля упругости при растяжении*

|  |  |  |
| --- | --- | --- |
| Модель | MSE | R2 |
| Linear Regression | 3.5310902339377382e-31 | 1.0 |
| KNeighborsRegressor | 3.5310902339377382e-31 | 1.0 |
| Decision Tree Regressor | 0.002268699766192818 | 0.9207559947003582 |

**2.4. Нейронная сеть для соотношения «матрица – наполнитель»**

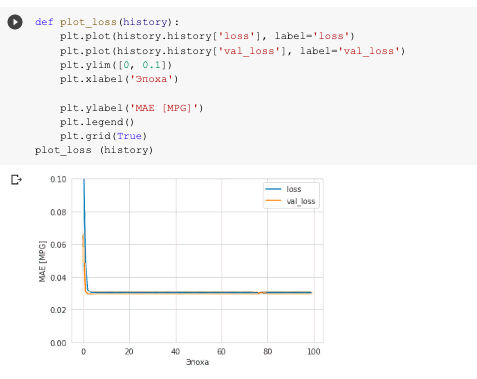
Разделение на тестовую и обучающую выборки осущестлялось также по принципу 70/30 (test\_size равен 0.3).

Построение нейросети:



Количество оптимизируемых параметров равно 34059.

Ошибки и потери:



MSE: 0.03168300828705616

R2: -0.00039930547597832877.

**Заключение**

На основании проведенного анализа можно предположить несостоятельность выбранных моделей для прогнозирования, запрошенное условиями задания.

Применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов. Так, коэффициент детерминации должен принимать значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Но при расчете прочности при растяжении R2 в лучшем случае ближе к 0, а для дерева решений принял отрицательное значение. А при оценке моделей для прогноза значений модуля упругости при растяжении высокой оказались значения средней квадратической ошибки. По тем же соображениям считаем малоуспешной реализацию нейронной сети.

Все же укажем, что лучшие результаты из предложенных моделей для модуля упругости при растяжении продемонстрировала линейная регрессия и метод ближайших соседей, для прочности при растяжении – дерево решений.

Также отметим, что в последующих этапах изучения предоставленных материалов автором предполагается применение большего числа моделей, в том числе и не относящихся к регрессиям, удаление выявленных выбросов, преобразование категориальных признаков и другие манипуляции для большей эффективности исследования.

**Список используемых источников**

1. Абрамов Р. Что такое дерево решений и где его используют? [Электронный ресурс]: https://habr.com/ru/company/productstar/blog/523044/

2. Григорьев Е.А., Климов Н.C., Разведочный анализ данных с помощью Python // E-Scio, 2020

3. Давыдова И.Ф., Кавун Н.С., Швецов Е.П. Базальтопластики для работ при повышенных

4. Макин А. Сумма технологии. Как композитные материалы остаются «на хайпе» более 5 тысяч лет [Электронный ресурс]: https://habr.com/ru/company/redmadrobot/blog/563336/

5. Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. СПб: Питер, 2013.

6. температурах // «Все материалы. Энциклопедический справочник», №6, 2012 г. [Электронный ресурс]: www.viam.ru/public

7. Чубукова, И. А. Data Mining. Интуит, 2016. [Электронный ресурс]: https://www.studentlibrary.ru/book/ISBN9785947748192.html

1. Макин А. Сумма технологии. Как композитные материалы остаются «на хайпе» более 5 тысяч лет [Электронный ресурс]: https://habr.com/ru/company/redmadrobot/blog/563336/ [↑](#footnote-ref-1)
2. Там же [↑](#footnote-ref-2)
3. Давыдова И.Ф., Кавун Н.С., Швецов Е.П. Базальтопластики для работ при повышенных

   температурах // «Все материалы. Энциклопедический справочник», №6, 2012 г. [Электронный ресурс]: www.viam.ru/public [↑](#footnote-ref-3)
4. Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. СПб: Питер, 2013. [↑](#footnote-ref-4)
5. Чубукова, И. А. Data Mining. Интуит, 2016. [Электронный ресурс]: <https://www.studentlibrary.ru/book/ISBN9785947748192.html> [↑](#footnote-ref-5)
6. Абрамов Р. Что такое дерево решений и где его используют? [Электронный ресурс]: <https://habr.com/ru/company/productstar/blog/523044/> [↑](#footnote-ref-6)
7. Григорьев Е.А., Климов Н.C., Разведочный анализ данных с помощью Python // E-Scio, 2020 [↑](#footnote-ref-7)