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»

**Прогнозирование конечных свойств**

**новых материалов (композиционных материалов**)

Слушатель В.С. Благодатских

Москва, 2022

**Содержание**

[Введение 3](#_Toc106310043)

[1. Постановка задачи, описание исходных данных 4](#_Toc106310044)

[2. Разведочный анализ данных и визуализация исходных данных 5](#_Toc106310045)

[3. Предобработка данных 10](#_Toc106310046)

[4. Обучение моделей для прогноза модуля упругости при растяжении и прочности при растяжении. 11](#_Toc106310047)

[4.1. Разбиение датасета для дальнейшего обучения моделей. 11](#_Toc106310048)

[4.2. Обучение моделей линейной регрессии. 13](#_Toc106310049)

[4.3. Обучение модели К-ближайших соседей (KNeighborsRegressor) 17](#_Toc106310050)

[4.4. Обучение модели «Случайного леса» (Random Forest Regressor). Подбор параметров. 18](#_Toc106310051)

[5. Разработка нейронных сетей 21](#_Toc106310052)

[5.1. Нейронные сети с одним выходом для параметра соотношение матрица-наполнитель. 21](#_Toc106310053)

[5.2. Построение нейронной сети с один выходом для предсказания модуля упругости при растяжении 30](#_Toc106310054)

[6. Создание приложения с графическим интерфейсом для предсказания модуля упругости 33](#_Toc106310055)

[7. Заключение 34](#_Toc106310056)

[Список литературы 35](#_Toc106310057)

# Введение

В рамках выпускной квалификационной работы по курсу «Data Science» на тему «Прогнозирование конечных свойств новых материалов (композиционных материалов)» проведены исследования и анализ предоставленных данных о начальных свойствах компонентов композиционных материалов с использованием методов, изученных на курсе «Data Science».

Композиционный материал, композит — многокомпонентный материал, изготовленный (человеком или природой) из двух или более компонентов с существенно различными физическими и/или химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов и не являющимися простой их [суперпозицией](https://yandex.ru/q/question/obiasnite_prostymi_slovami_chto_takoe_9e9d5016/). При этом отдельные компоненты остаются таковыми в структуре композитов, отличая их от смесей и твёрдых растворов. В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.

1. **Постановка задачи, описание исходных данных**

Целью данной работы является обработка данных и построение при помощи методов машинного обучения модели прогнозирования характеристик «модуль упругости при растяжении» и «прочность при растяжении», рекомендации «соотношение матрица-наполнитель».

Исходные данные свойств композиционных материалов получены от структурного подразделения МГТУ им. Н.Э. Баумана Центра НТИ «Цифровое материаловедение: новые материалы и вещества».

Данные представляют собой характеристики и свойства композитных материалов, содержащиеся в двух таблицах Excel.

Исходные данные включают в себя:

1. Таблицу X\_bp.xlsx с количеством строк 1023, где находятся три исследуемых параметра (характеристики композита), семь входных (химико-физические соотношения и свойства используемых компонент), одна колонка с индексом (Рисунок 1).

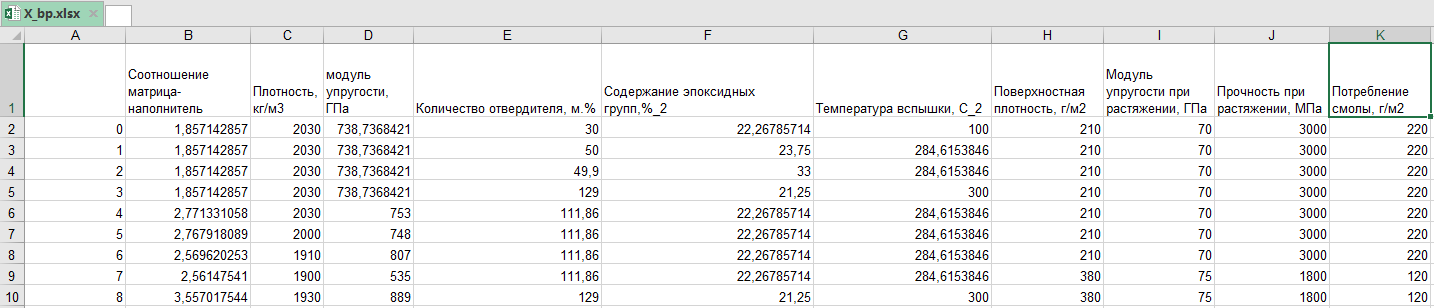


Рисунок 1 – Таблица X\_bp.xlsx

1. Таблицу X\_nup.xlsx с количеством элементов 1040, где находится три входных параметра (физические характеристики исследуемых образцов) (Рисунок 2) и одна колонка с индексом.

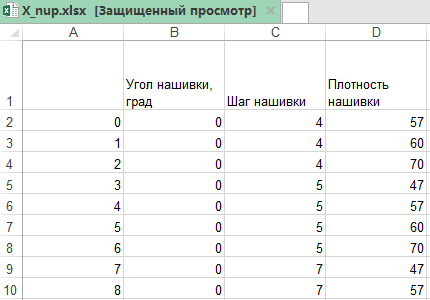


Рисунок 1 – Таблица X\_nup.xlsx

Колонка «А» с индексом в обеих таблицах служит для их объединения.

Анализ, предобработка данных, построение моделей выполнены посредством языка программирования Python, с применением библиотек numpy, pandas, matplotlib, seaborn, sklearn.

1. **Разведочный анализ данных и визуализация исходных данных**

Разведочный анализ данных в рамках поставленной задачи необходимо провести способом объединения по типу inner по полю индекса таблиц исходных данных

17 строк данных не были включены в общий датасет, так как они не имеют отличных строк в таблице X\_bp.xlsx

Сформированный исходный датафрейм содержит 1023 записи с входным параметрами и 3 выходными параметрами, пропуски значений отсутствуют, дублирующих значений нет (Рисунок 3).

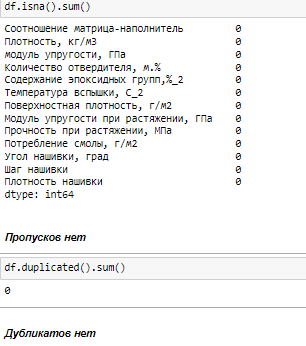


Рисунок 3 – Количество пропусков и дубликатов

Значения колонки 'Угол нашивки, град' были перекодированы в бинарный вид (Рисунок 4).

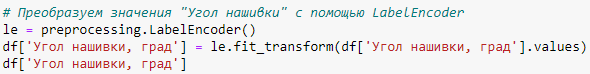


Рисунок 4 – Преобразование значений «Угол нашивки»

Описательная статистика (Рисунок 5), гистограммы распределения (Рисунок 6), диаграммы размаха «Ящик с усами» (Рисунок 7), позволяют получить наглядное представление о характерах распределений переменных.

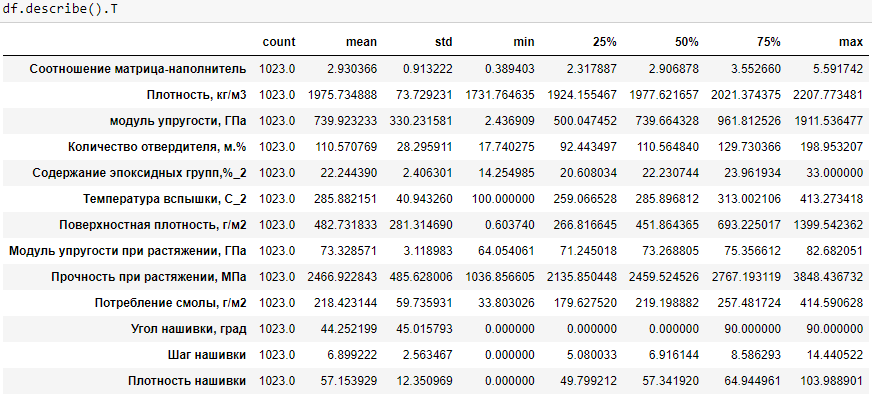


Рисунок 5 – Описательная статистика

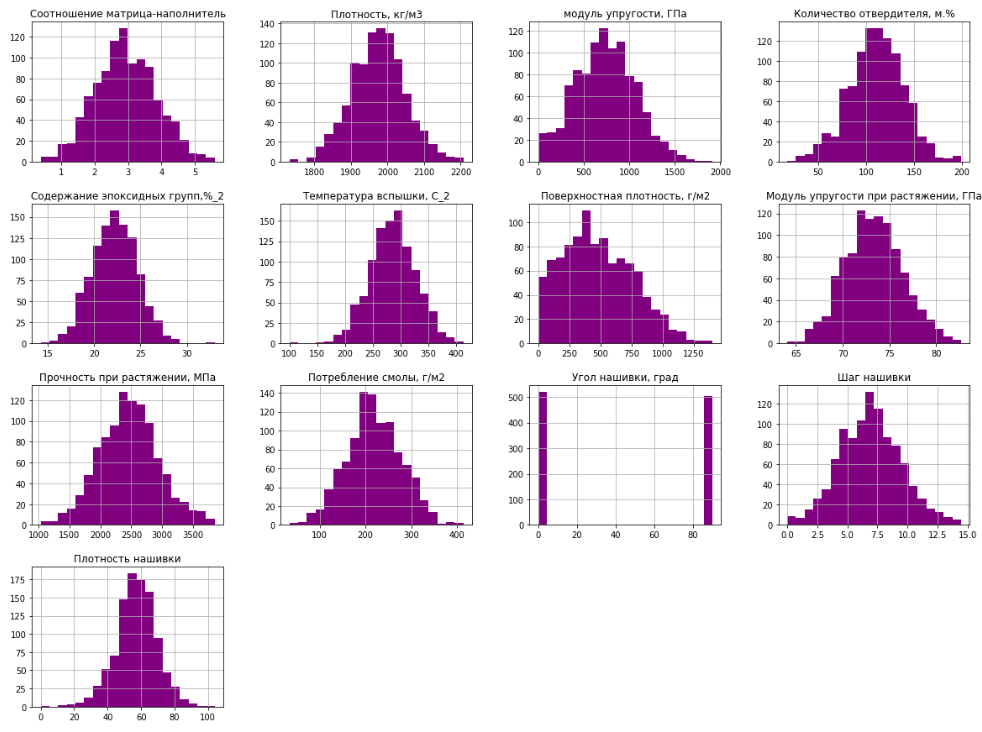


Рисунок 6 – Гистограммы распределения переменных

Из всех параметров, помимо «Угла нашивки», имеющего всего два значения, выделяются «Поверхностная плотность, г/м2» и «модуль упругости, ГПа» форма распределения менее других походит на нормальное.

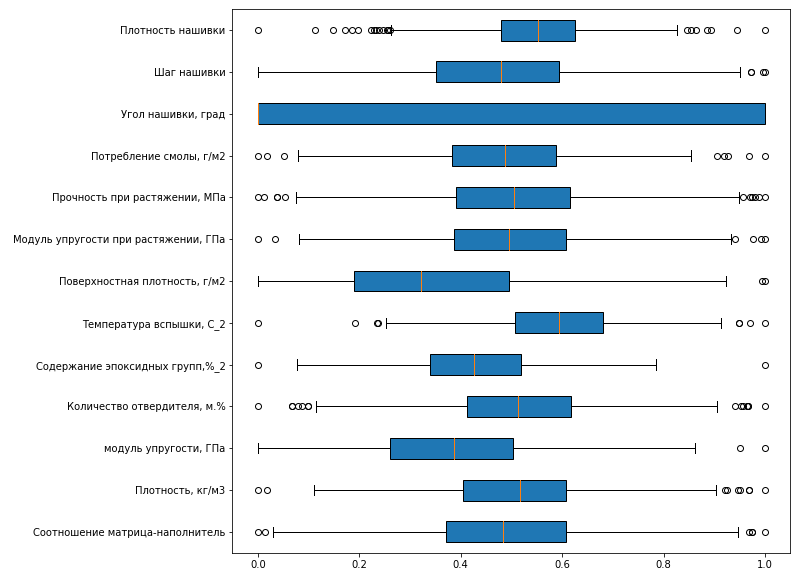


Рисунок 7 – Диаграммы размаха «Ящик с усами»

По диаграммам размаха видно, что по каждой характеристике, кроме «Угол нашивки, град», имеется наличие значений, находящихся за пределами полутора межквартильных расстояний от первого и третьего квартилей. Избавиться от них нам поможет «правило трех сигм» (Рисунок 8). Посчитаем и удалим выбросы (Рисунок 9).

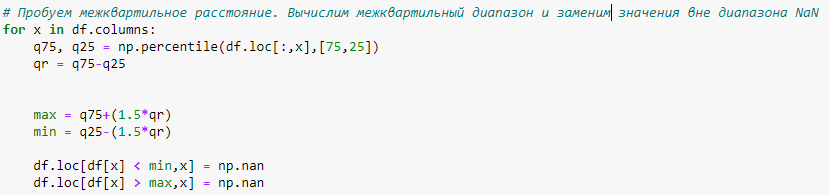


Рисунок 8 – Используем «правило трех сигм»

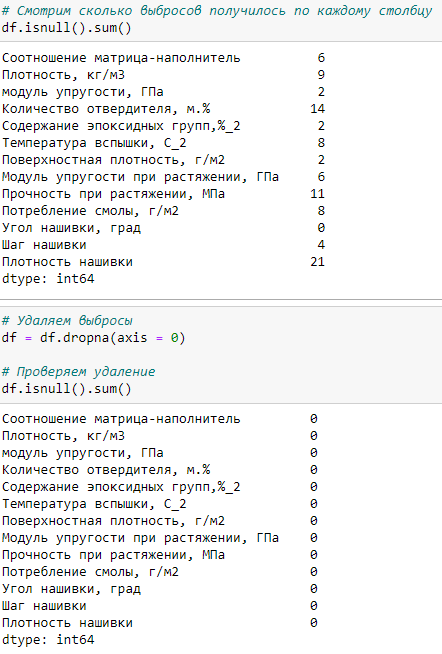


Рисунок 9 – Расчет и удаление выбросов

Построим матрицу корреляции (Рисунок 10), посмотрим, какие линейные связи есть между не стандартизированными и не нормализированными данными

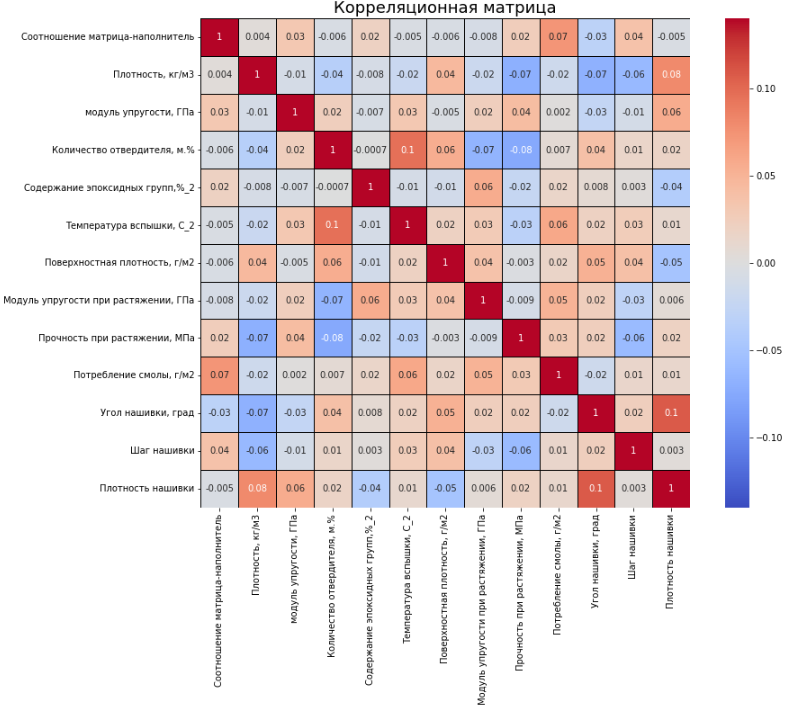


Рисунок 10 – Матрица корреляции

Видим очень слабую линейную зависимость. Максимальная положительная связь (0,1) есть между углом нашивки и поверхностной плотностью, что скорее всего, если угол нашивки равен , то поверхностная плотность ниже, если угол нашивки градусов, то поверхностная плотность выше.

Также есть положительная связь 0,1 между количеством отвердителя и температурой вспышки, т.е. если температура выше, то количество отвердителя выше.

Строим матрицу рассеяния scatter\_matrix (Рисунок 11)

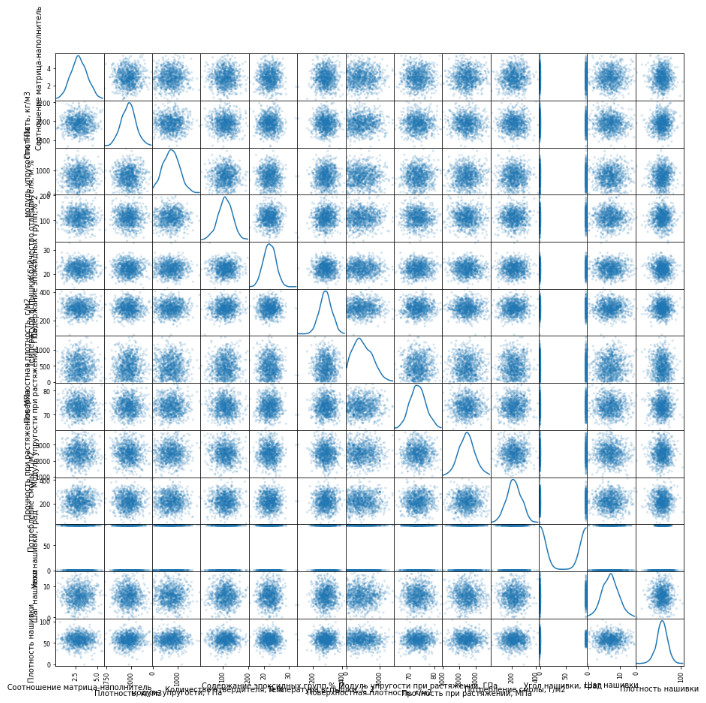


Рисунок 11 – Матрица рассеяния

Попарные графики рассеяния также не показывают линейной взаимосвязи между данными.

1. **Предобработка данных**

В ходе анализа данных было выполнено удаление выбросов, произведен поиск пропусков и дубликатов, на данном этапе проведем нормализацию и стандартизацию данных.

Выполним нормализацию с помощью метода MinMaxScaler. (Рисунок 12).

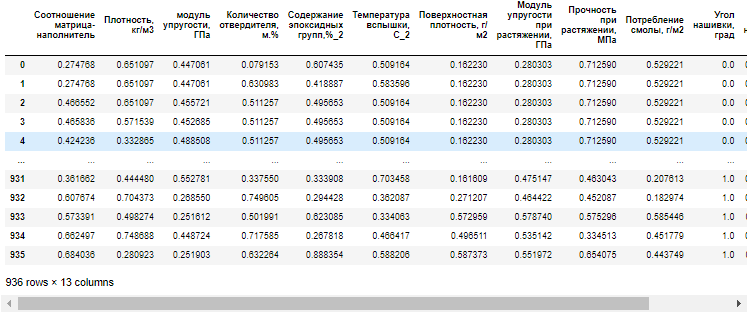


Рисунок 12 – Нормализация данных

Выполним стандартизацию данных с помощью StandardScaler (Рисунок 13)



Рисунок 13 – Стандартизация данных

После анализа и предобработки данных получаем датасет df\_standart с 936 уникальными строками.

1. **Обучение моделей для прогноза модуля упругости при растяжении и прочности при растяжении.**
   1. **Разбиение датасета для дальнейшего обучения моделей.**

Для обучения моделей прогноза параметров разобьем ранее подготовленный датасет следующим образом:

* Удаляем столбцы "Модуль упругости при растяжении" и "Прочность при растяжении". Назовем новый датасет "mu" и "pr".
* Запишем переменную "Модуль упругости при растяжении" в "mu\_Y", а датасет "mu" с удаленными столбцами в "mu\_X".
* Запишем переменную "Прочность при растяжении" в "pr\_Y", а датасет "pr" с удаленными столбцами в "pr\_X".

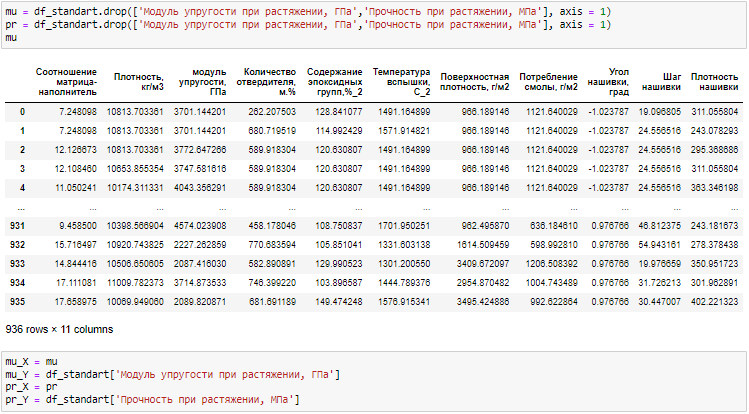


Рисунок 14 – Разбивка датасета и запись в переменные

Разобьем датасет mu\_X и датасет pr\_X на тестовую и тренировочную выборки, 30 процентов данных оставим на тестирование модели, на остальных происходит обучение моделей.





Рисунок 15 – Разбивка датасетов на тренировочную и обучающую выборки

Пишем функцию, которая рассчитывает среднее значение по тестовой выборке. Со средним будем сравнивать результаты предсказаний моделей (Рисунок 16)

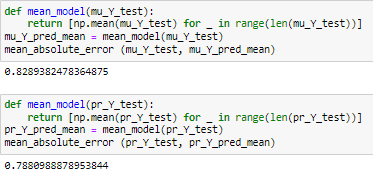


Рисунок 16 – MAE для модели среднего значения.

* 1. **Обучение моделей линейной регрессии.**

Построим модель линейной регрессии для модуля упругости при растяжении (Рисунок 17)

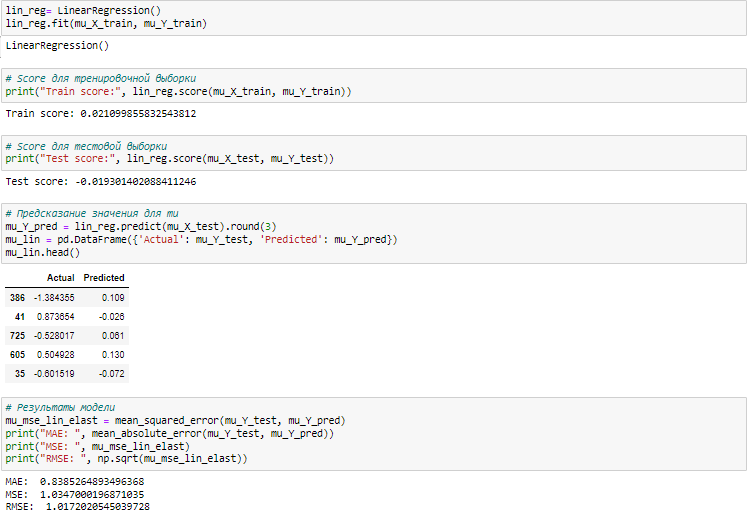


Рисунок 17 - Модель линейной регрессии для модуля упругости при растяжении



Рисунок 18 – Разброс точек, предсказанных и фактических данных

Видим, что разброс у фактических данных большой – линейная регрессия не очень хорошо прогнозирует зависимости.



Рисунок 18 – R2-score

Для линейной регрессии – MSE 1.035, R-score -0.019. Для модели, что использует среднее, MAE – 0.83, что говорит, что линейная регрессия плохо предсказывает поведение зависимой переменной от входных данных.

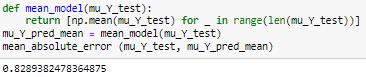


Рисунок 19 – МАЕ средняя

Построим линейную регрессию для Прочности при растяжении (Рисунок 20).

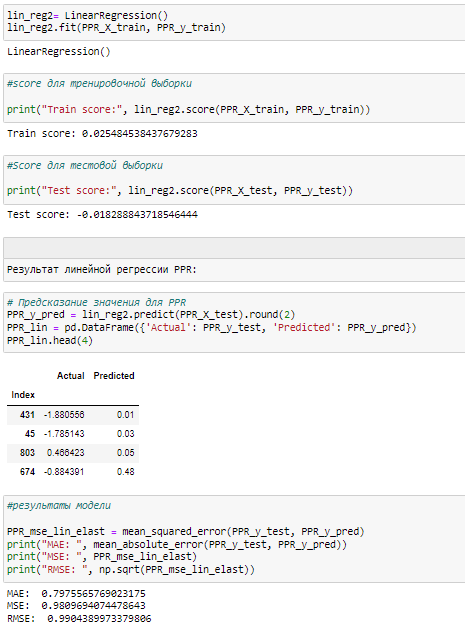


Рисунок 20 – Модель линейной регрессии для прочности при растяжении



Рисунок 21 – Разброс точек фактических и прогнозных данных

Видим, что разброс у фактических данных большой – линейная регрессия не очень хорошо прогнозирует зависимости.



Рисунок 22 – R2-score

Для линейной регрессии PPR - MAE 0.98, R-score -0.018. Для модели, что использует среднее, MAE – 0.79, что говорит, что линейная регрессия плохо предсказывает поведение зависимой переменной от входных данных.

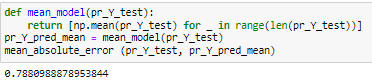


Рисунок 23 – МАЕ средняя

* 1. **Обучение модели К-ближайших соседей (KNeighborsRegressor)**

Построим модель К-ближайших соседей для модуля упругости при растяжении (Рисунок 24).



Рисунок 24 – Модель К-ближайших соседей для модуля упругости при растяжении

Построим модель К-ближайших соседей для прочности при растяжении (Рисунок 25).

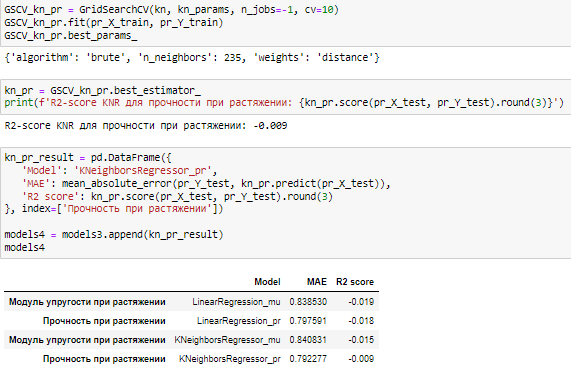
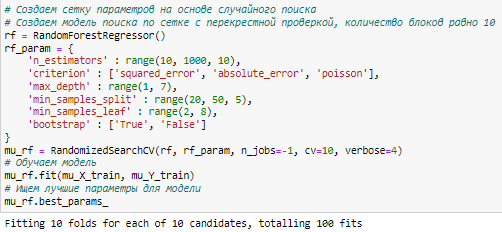
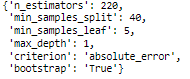


Рисунок 25 – Модель К-ближайших соседей для прочности при растяжении

Если коэффициент детерминации R2 – score равен нулю, то это говорит о том, что модель прогнозирует с таким же результатом, что, если бы мы взяли среднее значение прогнозируемого параметра. Если R2 – score меньше 0, это значит, что разработанная модель дает прогноз хуже, чем обычное усреднение.

* 1. **Обучение модели «Случайного леса» (Random Forest Regressor). Подбор параметров.**

Построим модель случайного леса для модуля упругости при растяжении (Рисунок 26).

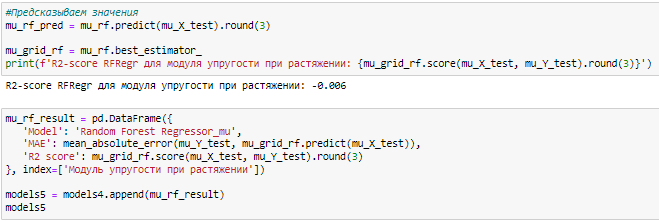
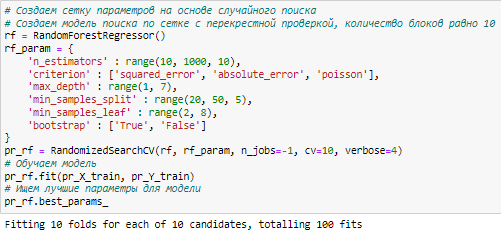
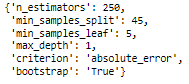


Рисунок 26 – Модель «Случайного леса» для модуля упругости при растяжении

Построим модель случайного леса для прочности при растяжении (Рисунок 27).

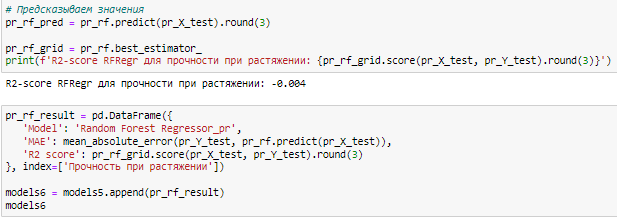


Рисунок 27 – Модель «Случайного леса» для прочности при растяжении

Все простроенные модели дают прогноз хуже, чем обычное усреднение (Рисунок 27).



Рисунок 28 – Сравнение моделей

Построенные модели прогнозирования не демонстрируют приемлемого уровня точности, и свидетельствуют о слабой линейной связи между данными датасета. Об этом на этапе разведывательного анализа данных говорит слабая попарная корреляция выходных данных с входными (матрица корреляции). Однако разведывательный анализ данных и теоретическая база композитных материалов также позволяет делать допущение, что связь возможна, но для ее обнаружения необходимо использовать более сложные способы исследования.

Также можно провести дополнительные способы разбивки данных на более малые выборки, уменьшение размерности. Разведочный анализ данных показывает, что Угол нашивки равен 0 градусов (520 наблюдений исходной выборки), а другая часть Угол нашивки равен 90 градусов (520 наблюдений исходной выборки). Так как данный входной параметр четко делится на две подгруппы, то имеет смысли, опираясь на теорию, рассмотреть модели для каждой из этих подвыборок. Максимальная положительная связь (0,1) есть между Углом нашивки и Поверхностной плотностью, что скорее всего, если Угол нашивки 0, то Поверхностная плотность ниже, если Угол нашивки 90 градусов, то Поверхностная плотность выше. Также есть положительная связь (0,1) между Количеством отвердителя и Температурой вспышки, т.е. если температура выше, значит, и Количество отвердителя выше. В теории, если Температура вспышки выше, то, нужно большее Количество отвердителя (как канифоль). Ничего нового, кроме обычной физики, данная связь не показывает.

Кроме того, теоретические исследования в области композитных материалов позволяют предположить, что не хватает дополнительных параметров (например, коэффициент Пуассона или данных об угле и направлении, под которыми располагают нити накладываемых слоёв), что также влияет на прочность при растяжении и модуль упругости при растяжении.

Из аналогичных исследований можно взять для тестирования подход решения задач моделирования композиционного материала с заданными электрофизическим свойствами (композиты на основе нанопористого кремния), на основе генетического алгоритма или рекуррентного типа нейронной сети(диэлектрическая проницаемость, проводимость)**.**

1. **Разработка нейронных сетей**
   1. **Нейронные сети с одним выходом для параметра соотношение матрица-наполнитель.**

Обобщающая способность (generalization) - способность сети работать на тех данных, что она не видела. Если нейронная сеть переобучена, то она адаптируется к данным, которые она видела (training set) и хорошо работает на них, но ее обобщающая способность снижается (плохо работает на данных, которые она не видела). Оценивают качество сети на тех, данных, которые она не видела (test set).

Если модель дает более стабильный результат, значит, данная модель является более лучшим решением.

Построим нейронную сеть с одним выходом для параметра соотношение матрица-наполнитель (Рисунок 29.1-29.3).

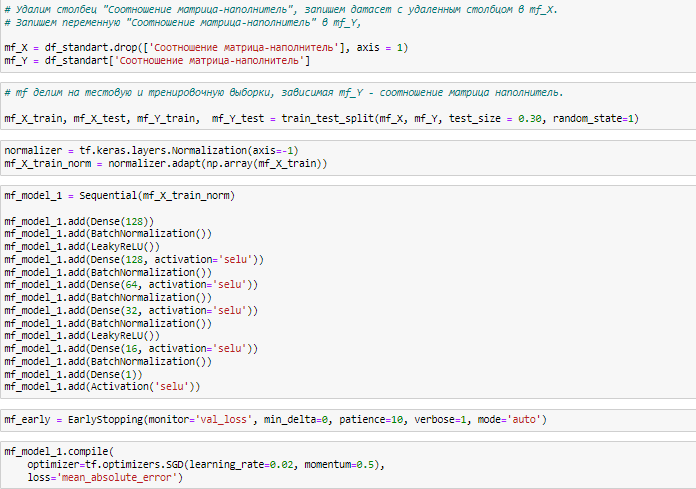


Рисунок 29.1 - Нейронная сеть с одним выходом для соотношения матрица-наполнитель (модель 1)

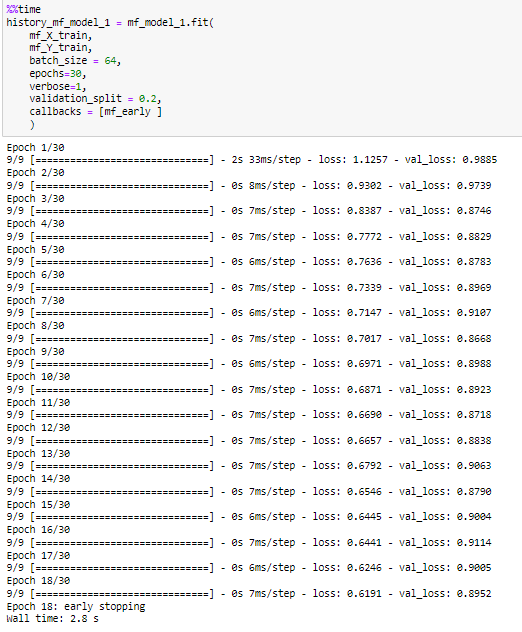


Рисунок 29.2 - Нейронная сеть с одним выходом для соотношения матрица-наполнитель (модель 1)

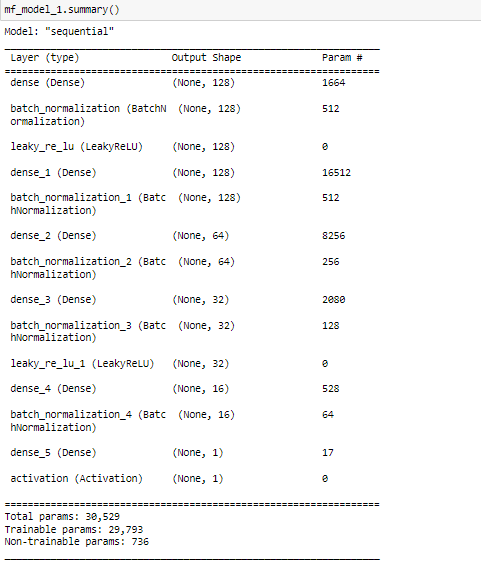


Рисунок 29.3 - Нейронная сеть с одним выходом для соотношения матрица-наполнитель (модель 1)

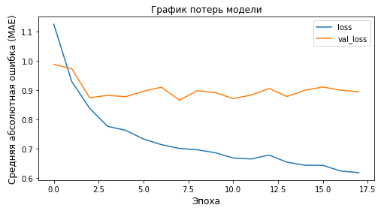


Рисунок 30 – График потерь модели 1

В примере видно (Рисунок 29-30), что после 3-й эпохи (val\_loss =0.8746) доля потерь на проверочной выборке начинает расти и к 13-й эпохе становится равной 0,963. Это означает, что модель переобучена и обучение необходимо останавливать после второй эпохи. И данная структура модели очень сложная для данного датасета.

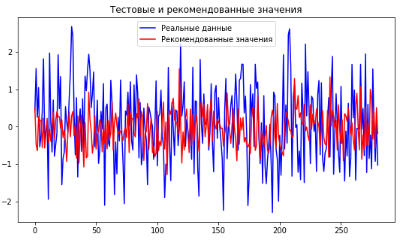


Рисунок 31 – Поведение тестовых и прогнозных значений

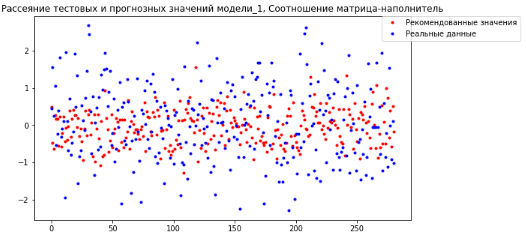
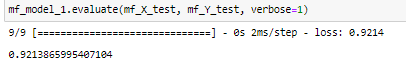


Рисунок 32 - Рассеяния тестовых и прогнозных значений

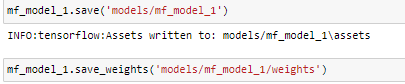
Метод evaluate () возвращает значения функции потерь и метрики для обученной модели.





Нейронная сеть (Модель 1) хуже (loss = 0.921), чем средняя предсказывает поведение зависимой переменной (MAE =0.83).

Сохраним модель



Построим нейронную сеть попроще с одним выходом для параметра соотношение матрица-наполнитель (Рисунок 33)

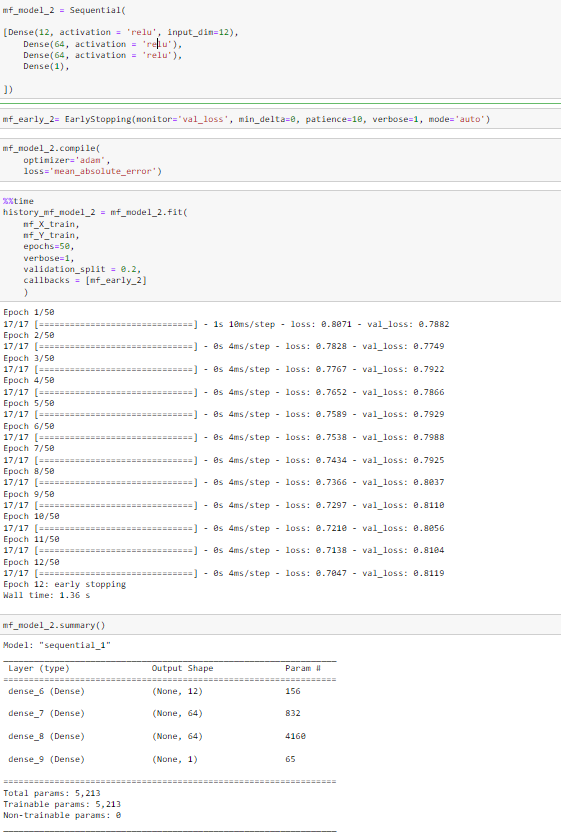


Рисунок 33 - Нейронная сеть с одним выходом для соотношения матрица-наполнитель (модель 2)

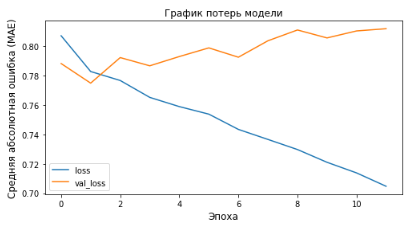


Рисунок 34 - График потерь модели 2

В примере видно, что после 4-й эпохи (val\_loss =0.786) доля потерь на проверочной выборке начинает расти и к 12-й эпохе становится равной 0,8119.

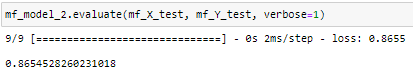
Это означает, что модель переобучена и обучение необходимо останавливать после четвертой эпохи. И данная структура модели очень сложная для данного датасета.



Рисунок 35 – Поведение тестовых и прогнозных значений



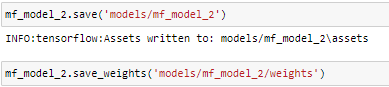
Рисунок 36 - Рассеяния тестовых и прогнозных значений





Нейронная сеть (Модель 2) чуть лучше (loss=0.87) предсказывает поведение зависимой переменной, чем (Модель 1), но снова хуже, чем обычная средняя (MAE =0.83).

Сохраним модель



Сравнение качества моделей нейронных сетей, решающих задачу регрессии, проводится на показателе MSE/ MAE (на тестовой выборке). Если порядок ошибки примерно одинаковый для нескольких моделей, выбирается та сеть, которая проще по архитектуре.

Если на тренировочном наборе данных ошибка снижается, а на валидационном растет, то значит, наша модель переобучилась.

Для нахождения более удачной модели необходимо менять гиперапараметры обучения сети, а именно:

* Количество нейронов входного слоя. Количество нейронов задается при создании входного слоя: model.add (Dense (XXX, …..)). Использовать разные значения. Выбрать значение, при котором самая высокая доля верных ответов нейросети на тестовых данных.
* Количество слоёв обучения (dance). Добавляем/ убираем скрытый слой. Можно добавить скрытый слой с разным количеством нейронов: 500, 700, 900 и 1200. Выбрать наиболее подходящее количество нейронов скрытого слоя. Оценить, как изменяется время обучения при добавлении/удалении скрытого слоя с разным количеством нейронов. В идеале, можно посмотреть сеть минимального размера, что может дать лучший результат.
* Количество эпох обучения (epochs). Можно обучать сеть в течение 50, 75, 100 и 125 эпох. Выбрать количество эпох, при котором самая высокая доля верных ответов нейросети на тестовых данных.
* Размер мини-выборки (batch\_size).Использовать размер мини-выборки 50, 100, 200 и 400. Выбрать значение, при котором самая высокая доля верных ответов нейросети на тестовых данных.
* Размер параметра скорости обучения. Он важен при углублении структуры сети.
* Архитектура нейронной сети; поменять активационную функцию.

Чтобы создать сеть с лучшими значениями всех гиперпараметров обучения, необходимо вести запись всех версий модели с изменёнными гиперпараметрами, которые определили на предыдущем шаге, и отслеживать увеличилась ли доля верных ответов нейросети.

Полученные результаты могут сократить сроки и издержки получения новых композитов с заданными характеристиками.

* 1. **Построение нейронной сети с один выходом для предсказания модуля упругости при растяжении**

Построим нейронную сеть с одним выходом для модуля упругости при растяжении (Рисунок 37.1-37.2).

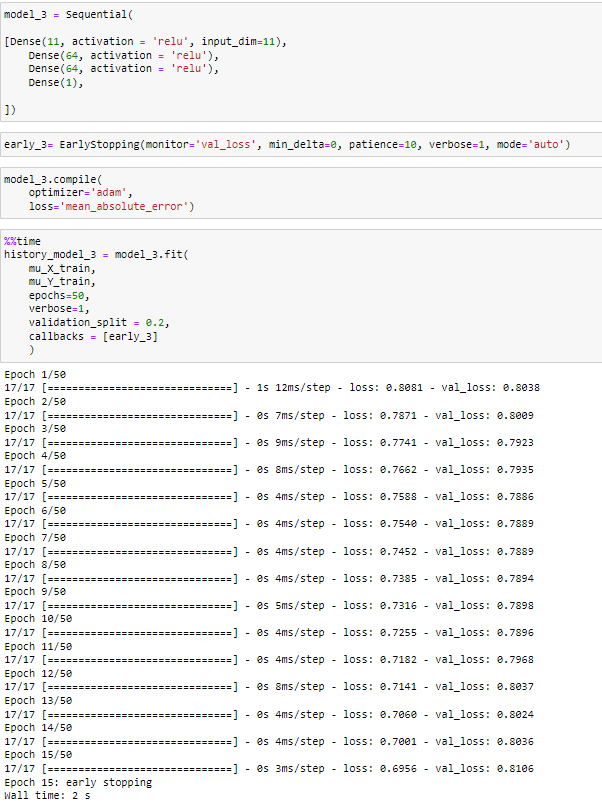


Рисунок 37.1 - Нейронная сеть с одним выходом для предсказания модуля упругости при растяжении

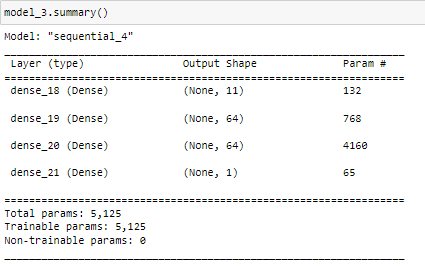


Рисунок 37.2 - Нейронная сеть с одним выходом для предсказания модуля упругости при растяжении

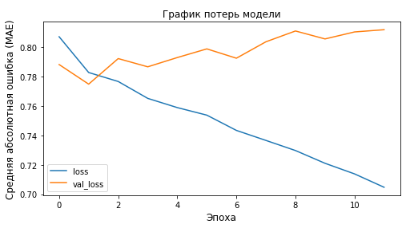


Рисунок 38 – График потерь модели



Рисунок 39 – Поведение тестовых и прогнозных значений

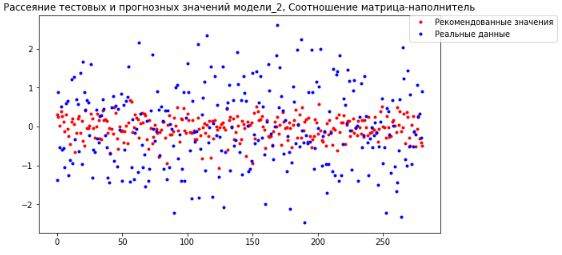
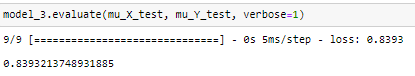


Рисунок 40 – Рассеяния тестовых и прогнозных значений





Нейронная сеть (loss=0.84) предсказывает поведение зависимой переменной хуже, чем обычная средняя (MAE =0.83), но лучше, чем предыдущие модели.

Сохраним модель.

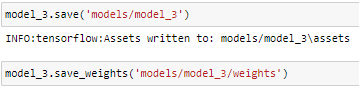


Рисунок 41 – Сохранение модели

1. **Создание приложения с графическим интерфейсом для предсказания модуля упругости**

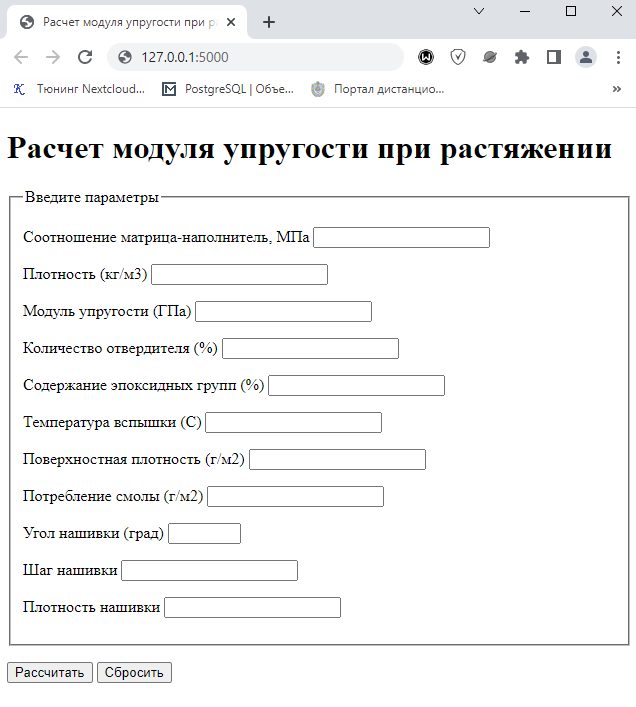


Рисунок 41 – Flask-приложение

Ссылка на репозиторий GitHub <https://github.com/Whale58/BMSTU>

1. **Заключение**

Данные и модели показывают, что при получении большего объема данных, их более детального изучения и обработки, можно значительно улучшить модели или применить другие для предсказаний искомых параметров.

На момент написания работы мне не хватило знаний и времени для получения более качественных моделей и крайне не хватило времени на более детальное изучение и предобработку данных.

Курс «Data Science» был для меня очень полезен и в очередной раз я убедился, что при желании и достаточном количестве времени возможно освоить новую профессию. Данное обучение позволило мне приобрести необходимые навыки и заинтересовать меня для продолжения развития в данном направлении. Некоторые навыки, приобретенные на данном курсе, уже частично помогают мне в работе.

**Список литературы**

1. Язык программирования Python – Режим доступа: https://www.python.org/. (дата обращения 16.04.2022)
2. Библиотека Matplotlib – Режим доступа: https://matplotlib.org/. (дата обращения 16.05.2022)
3. Д. Фостер Генеративное глубокое обучение. Творческий потенциал нейронных сетей. - СПб.: Питер. - 2020. - 336 с. - ISBN: 978-5-4461-1566-2
4. С. Николенко, А. Кадурин, Е. Архангельская Глубокое обучение. Погружение в мир нейронных сетей. - СПб.: Питер. - 2020. - 480 с. ISBN: 978-5-4461-1537-2
5. Андерсон, Карл Аналитическая культура. От сбора данных до бизнес-результатов / Карл Андерсон ; пер. с англ. Юлии Константиновой ; [науч. ред. Руслан Салахиев]. — М. : Манн, Иванов и Фербер, 2017. — 336 с
6. Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).
7. Аллен Б. Дауни – Основы Python. Научитесь думать как программист / Аллен Б. Дауни ; пер. с англ. С. Черникова ; [науч. ред. А. Родионов]. — Москва : Манн, Иванов и Фербер, 2021. — 304 с.
8. Библиотека Pandas – Режим доступа: https://pandas.pydata.org/. (дата обращения 16.05.2022)
9. Библиотека Sklearn- Режим доступа: https://scikit-learn.org/stable/. (дата обращения 16.5.2022)
10. Джулли, Пал: Библиотека Keras - инструмент глубокого обучения / пер. с англ. А. А. Слинкин.- ДМК Пресс, 2017. – 249 с.