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 Ремизова Светлана Николаевна

Москва, 2023

Содержание

Содержание………………………………………………………………………..

Введение…………………………………………………………………………..1

1.Аналитическая часть……………………………………………………………3

1.1.Постановка задачи…………………………………………………………….3

1.2.Описание используемых методов……………………………………………4

1.3.Разведочный анализ данных………………………………………………….7

2. Практическая часть…………………………………………………………...15

2.1.Предобработка данных. …………………………………………………….15

2.2. Разработка и обучение модели…………………………………………….20

2.3.Тестирование модели………………………………………………………..21

2.4.Написать нейронную сеть, которая будет рекомендовать соотношение

матрица …………………………………………………………………………..29

2.5.Разработка приложения……………………………………………………..33

2.6.Создание удаленного репозитория и загрузка результатов работы на

него……………………………………………………………………………….35

Список используемой литературы и веб русурсы……………………………….

# **Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

Композиционный материал или композит это многокомпонентный материал, изготовленные из двух или более компонентов с существенно- различными физическими или химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками.

В качестве наполнителей композитов, как правило выступают углеродные или стеклянные волокна, а роль матрицы играет полимер. Они состоят из двух или более компонентов, нерастворимых друг с другом, с чётко обозначенной границей и сильным взаимодействием по всей зоне контакта.

В составе композита принято выделять матрицу и наполнитель, последние выполняют функцию армирования. Матрица это непрерывная фаза, в которой нерастворимые материалы помещаются в другую природу, называемую арматурой или наполнителем.

Внедрение композиционных материалов обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства.

1

Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется.

У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить

характеристики композита, состоящего из этих компонентов, достаточно проблематично.

Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик.

Для достижения определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов и затраты на рабочую силу.

Целью данной работы является прогнозирование конечных свойств новых материалов на основе базальтопластика (композиционных материалов).

Суть прогнозирования заключается в моделировании репрезентативного элемента композитного объёма на основе данных о свойствах входящих компонентов компонентов (связующего и армирующего компонента).

Базальт - магматическая вулканическая порода. Это самая распространённая порода на поверхности Земли. Базальты образуются путём затвердевания силикатного магматического расплава. Активно развивается использование композитных материалов на основе базальта.

Базальтопластик - современный композитный материал на основе базальтовых волокон и органического связующего вещества. В настоящее время базальтопластик успешно конкурирует с металлическими изделиями, превосходя их по коррозионной и некоторым другим свойствам.

2

**1.Аналитическая часть**

**1.1.Постановка задачи**

Для исследовательской работы было дано 2 файла в формате Exsel c данными по композитным материалам.

В ходе работы необходимо изучить теоретические основы и методы решения поставленной задачи.

Провести разведочный анализ предложенных данных. Нарисовать гистограммы распределения, диаграммы ящика с усами, графики рассеяния точек. Получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков, предобработать данные: удалить шумы и выбросы, сделать нормализацию и стандартизацию.

Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении.

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

Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз соотношения «матрица-наполнитель».

Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

3

**1.2.Описание используемых методов**

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и традиционно это 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:

* случайный лес;
* линейная регрессия;
* К-ближайших соседей;
* дерево решений;

Случайный лес (RandomForest) — это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов.  Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив.

Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую масштабируемость.

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недо обучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

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.

4

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).

Достоинства метода: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.

Дерево принятия решений (DecisionTreeRegressor) – метод автоматического анализа больших массивов данных.  Это инструмент принятия решений, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево принятия решений - эффективный инструмент интеллектуального анализа данных и предсказательной аналитики.

Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область.

5

Регрессия дерева решений отслеживает особенности объекта и обучает модель в структуре дерева прогнозированию данных в будущем для получения значимого непрерывного вывода. Дерево решений один из вариантов решения регрессионной задачи, в случае если зависимость в данных не имеет очевидной корреляции.

Достоинства метода: помогают визуализировать процесс принятия решения и сделать правильный выбор в ситуациях, когда результаты одного решения влияют на результаты следующих решений; создаются по понятным правилам; просты в применении и интерпретации; заполняют пропуски в данных наиболее вероятным решением; работают с разными переменными; выделяют наиболее важные поля для прогнозирования;

Недостатки метода: ошибаются при классификации с большим количеством классов и небольшой обучающей выборкой; имеют нестабильный процесс (изменение в одном узле может привести к построению совсем другого дерева); имеет затратные вычисления; необходимо обращать внимание на размер; ограниченное число вариантов решения проблемы.

Коэффициент детерминации измеряет долю дисперсии, объяснённую моделью, в общей дисперсии целевой переменной.

Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

MSE (Mean Squared Error) или средняя квадратичная ошибка принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

6

**1.3. Разведочный анализ данных**

Цель разведочного анализа - получение первоначальных представлений о характерах распределений переменных исходного набора данных.

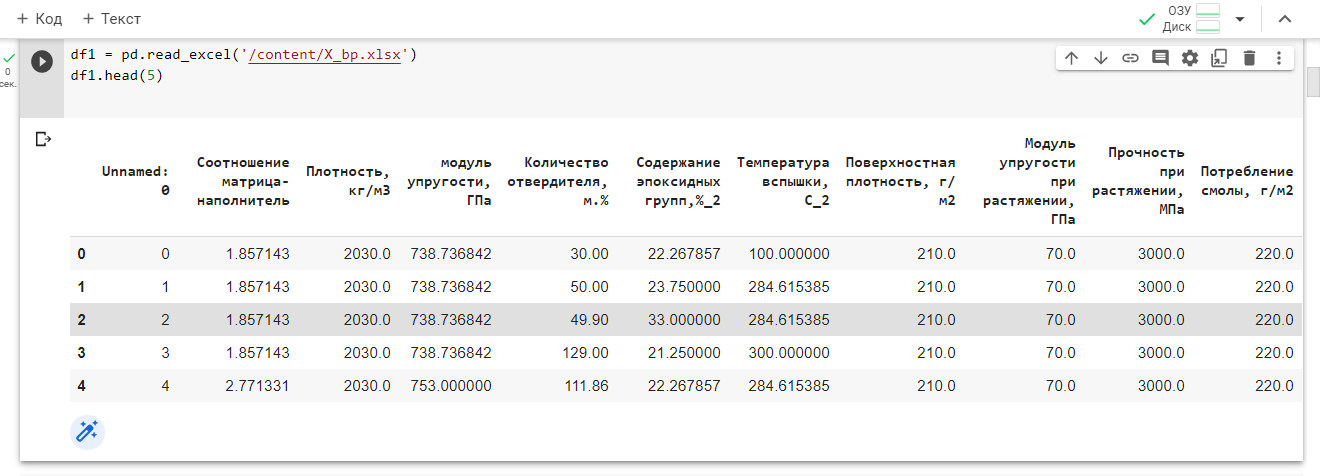
Формирование оценки качества исходных данных, выявление характера взаимосвязи между переменными, с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения, диаграммы ящика с усами, графики рассеяния точек, тепловая карта описательная статистика для каждой переменной, проверка наличия пропусков и дубликатов.

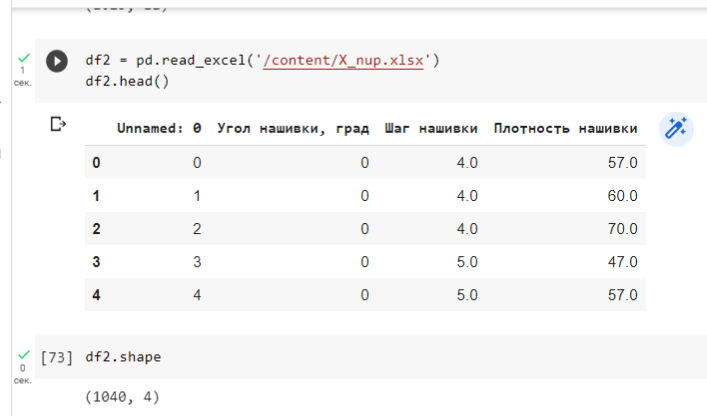
Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Очевидно, что необработанные данные могут содержать искажения и пропущенные значения – это может привести к крайне неверным результатам по итогам моделирования. Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

7

Для исследовательской работы было дано 2 файла: X\_bp.xlsxт, с состоящий из 1023 строки и 11 столбцов (рисунок 1) и X\_nup.xlsx , состоящий из 1040 строки и 4 столбцов (рисунок 2).



«Рисунок 1. Загрузка данных»

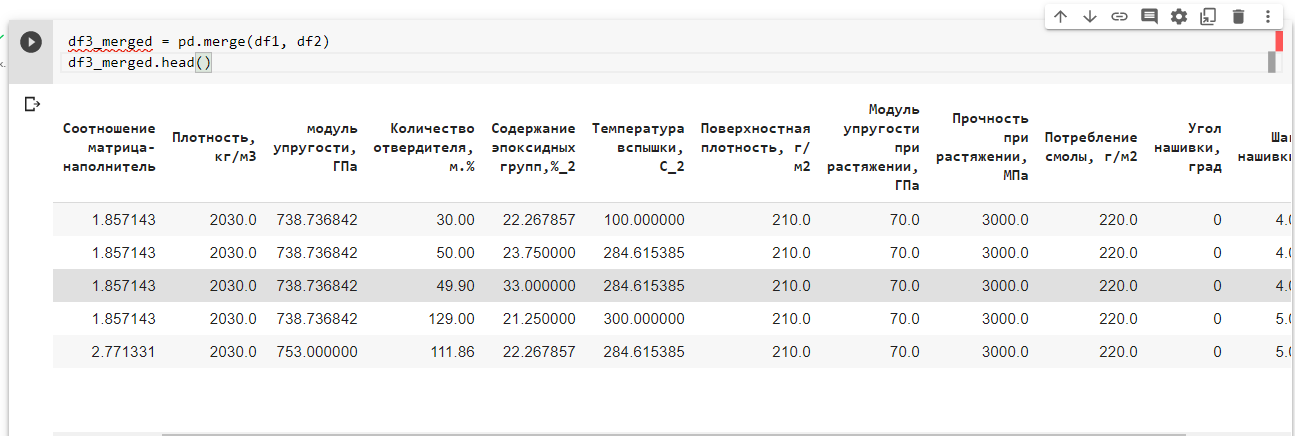


«Рисунок 2. Загрузка данных»

8

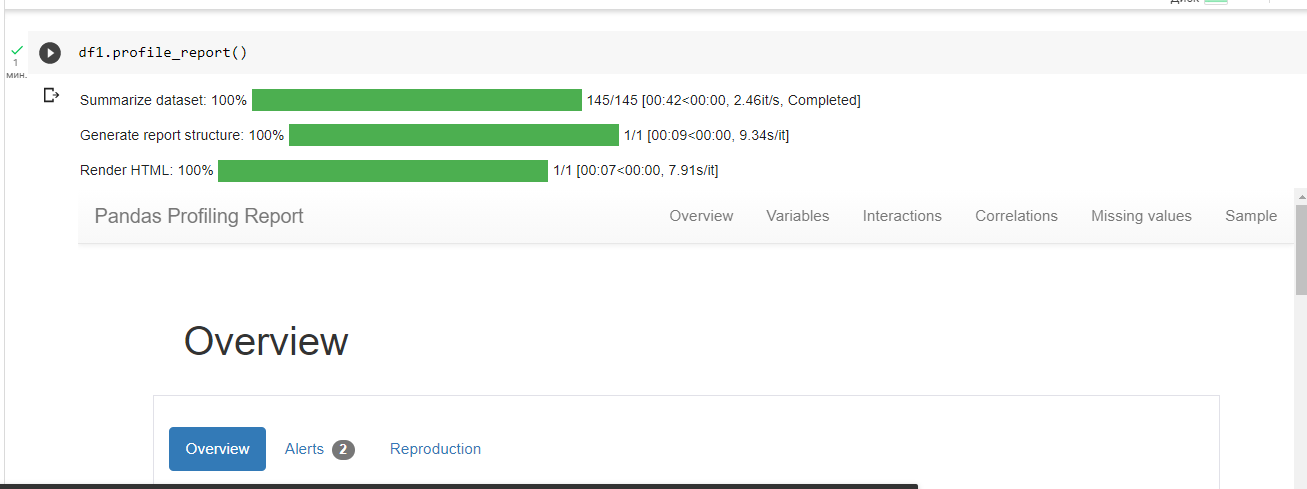
Цель работы разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель».

Для дальнейшей работы с данными нужно объединить 2 файла в одну таблицу. Фрагмент таблицы представлен ниже (рисунок 3).



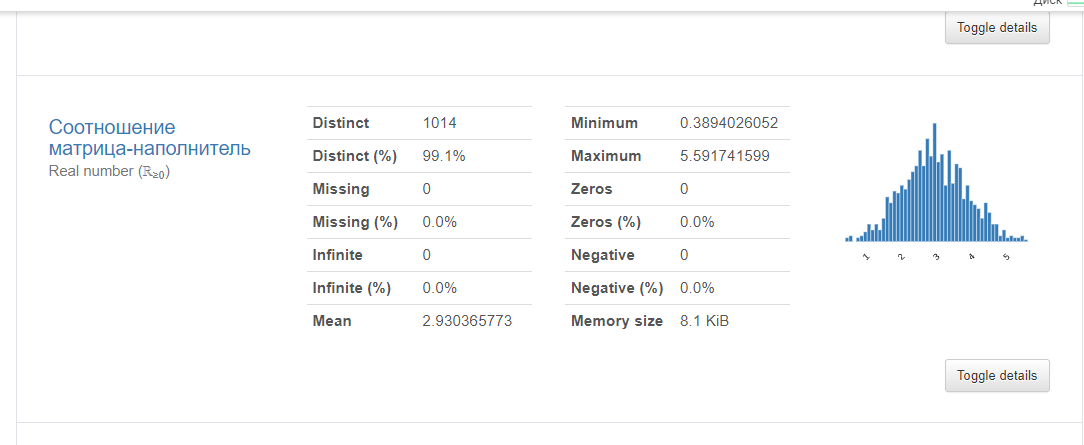
«Рисунок 3. Объединение двух таблиц»

Для быстрой, масштабной и наиболее полной информации по всем столбцам датасета, c применением различных визуализаторов, в данной работе был использован модуль библиотеки Pandas - profile\_reports( ), фрагменты использования представлены на рисунках 4,5, 6 и 7.



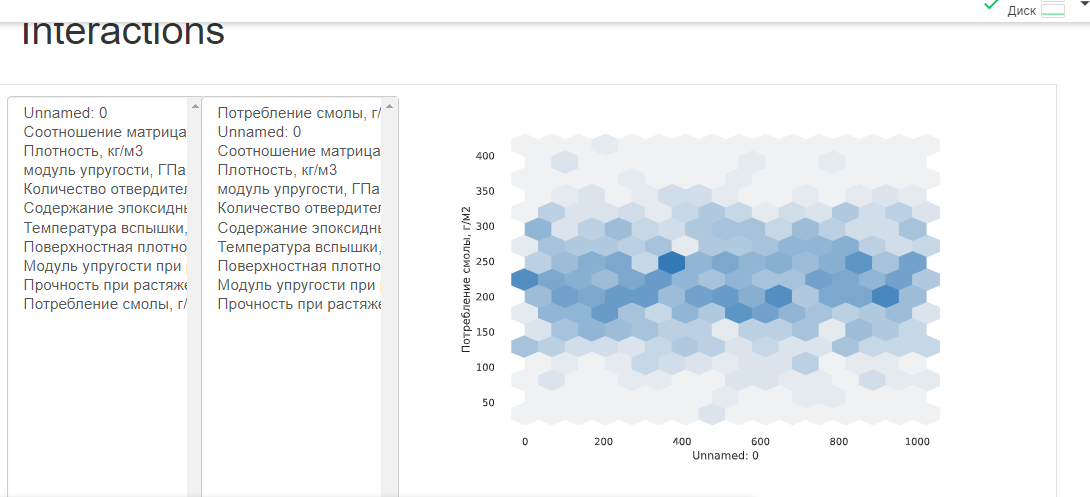
«Рисунок 4. Метод profile\_report»

9



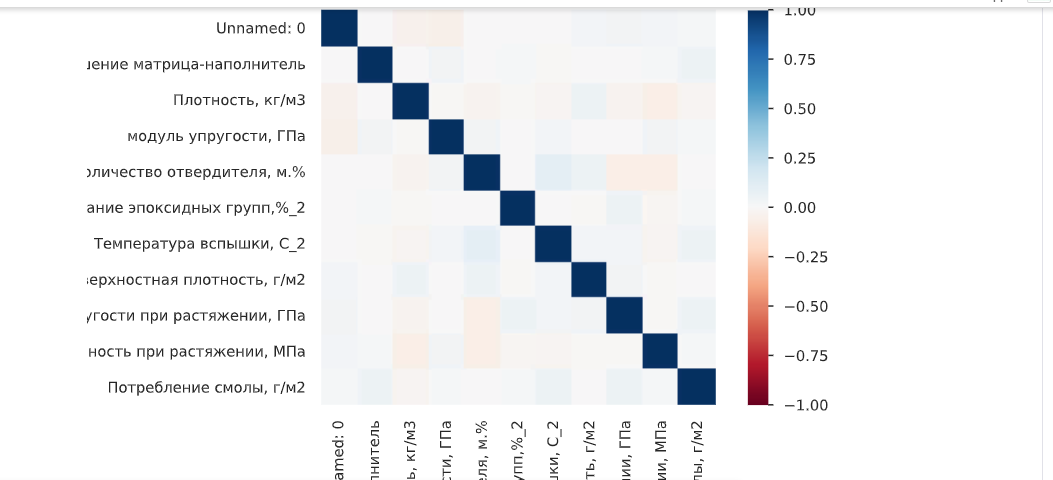
«Рисунок 5. Гистограмма распределения»

Автоматически, подобно представленному на рисунке 5, при помощи данного модуля, были выведены гистограммы распределения каждой переменной.



«Рисунок 6 Тепловая Матрица»

10

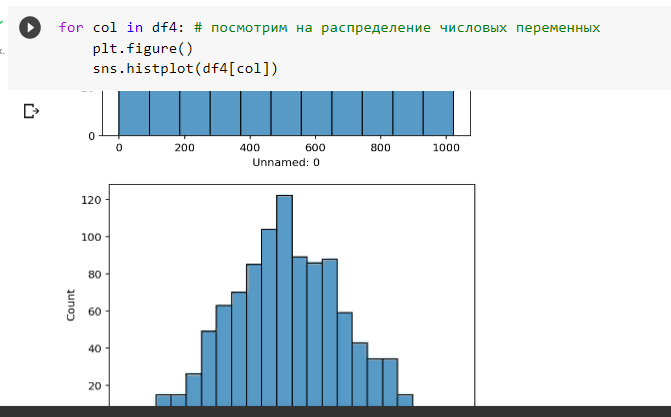


«Рисунок 7 Корреляционная матрица»

Проведен разведочный анализ данных, выведены гистограммы распределения числовых переменных. Фрагмент визуализации виден на рисунке 8, где можно увидеть, что данные имеют выбросы.

Гистограммы используются для изучения распределений частот значений переменных. Мы видим очень слабую корреляцию между переменными.

11

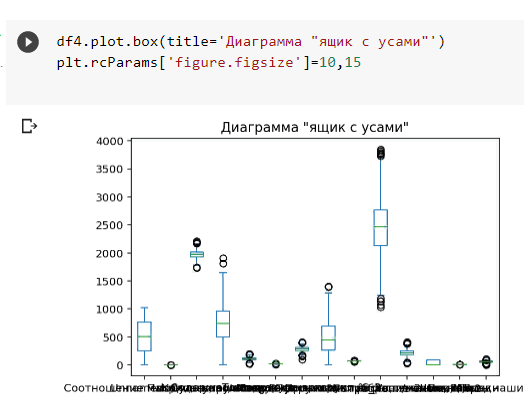


«Рисунок 8. Гистограмма распределения»

Далее были выведены диаграммы ящиков с усами и графики рассеяния

точек. Рисунок 9, 10 и 11. В фрагментах диаграмм ящиков с усами четко прослеживается наличие выбросов данных. Особенно это визуализируется на рисунке 10.

12

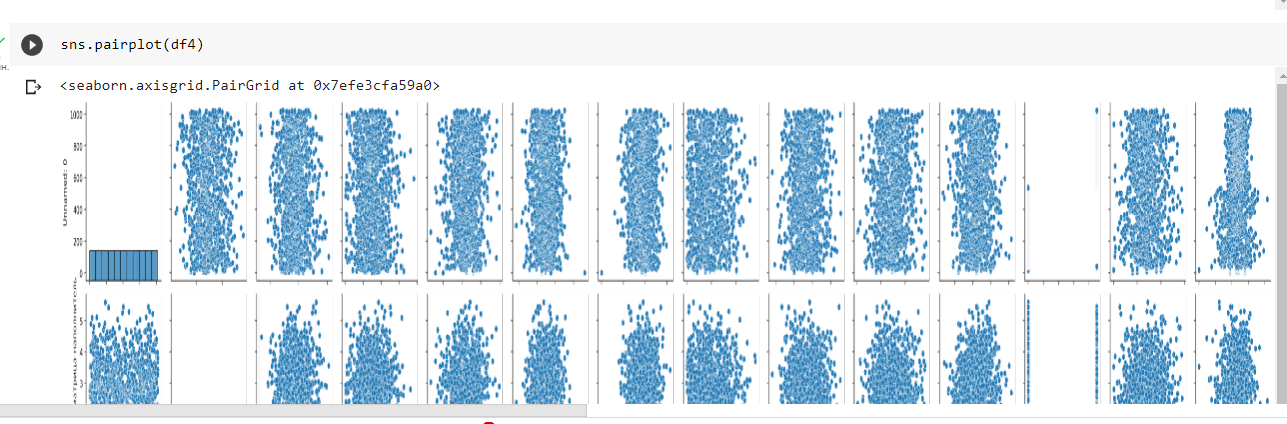


«Рисунок 9. Диаграмма ящик с усами»

13



«Рисунок 10.Диаграмма ящик с усами»



«Рисунок 11. График рассеяния точек»

14

**2. Практическая часть**

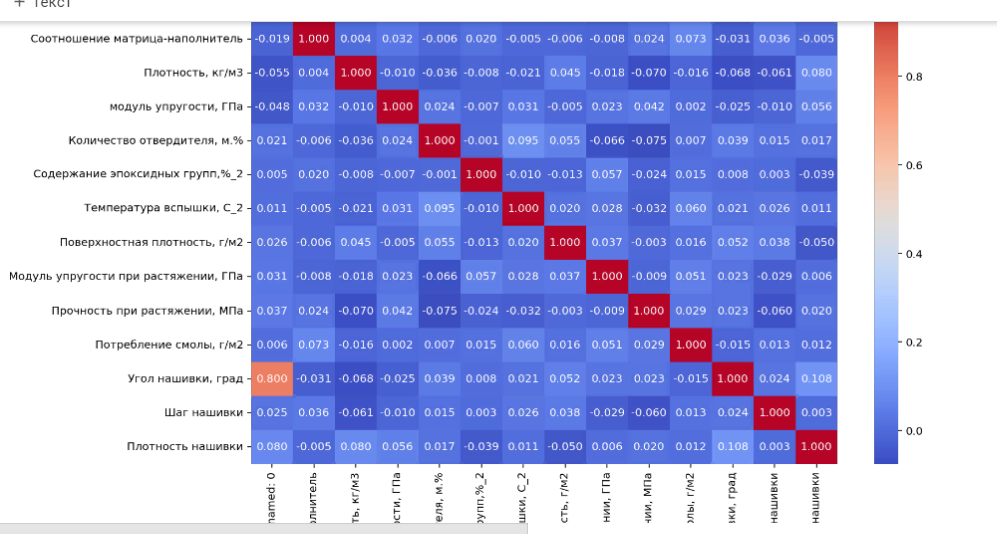
**2.1. Предобработка данных**.

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их.

Очевидно, что «грязные» и необработанные данные могут содержать искажения и пропущенные значения – это ненадёжно, поскольку способно привести к крайне неверным результатам по итогам моделирования.

Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

Данные объединённого датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции на рисунках 12 и 13

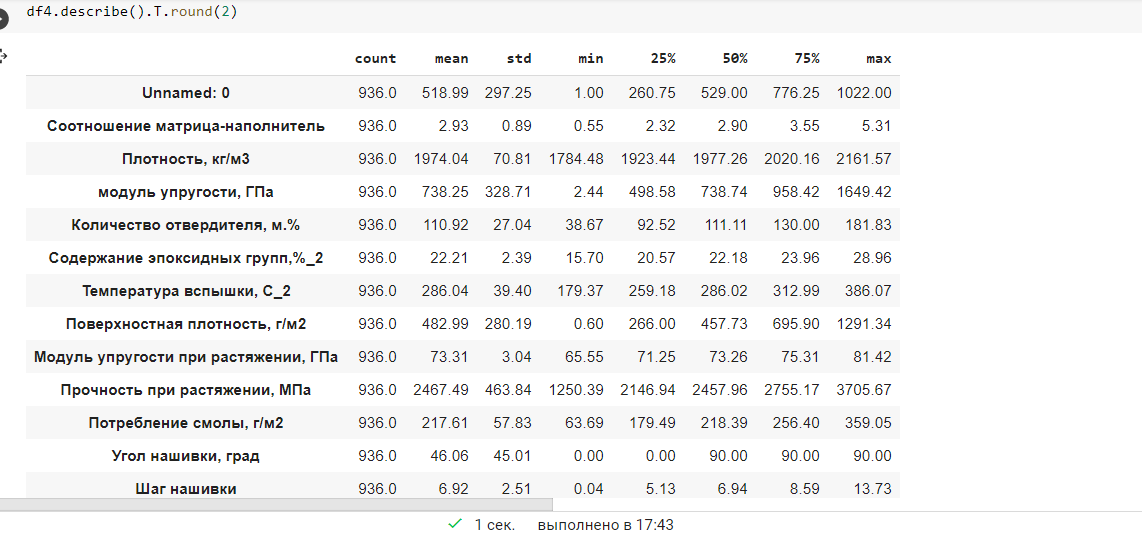


«Рисунок 12. Корреляционная матрица»

15



«Рисунок 13. Корреляционная матрица»



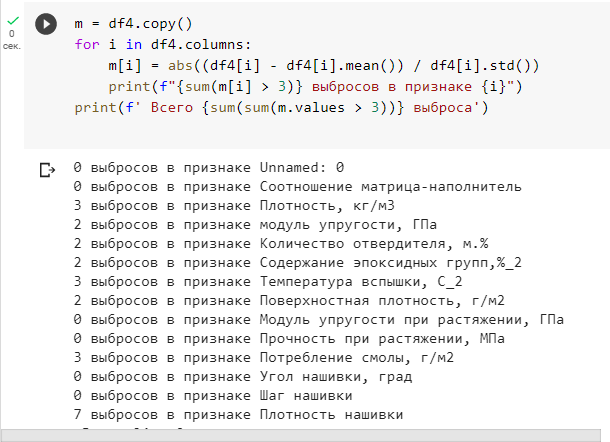
«Рисунок 14. Описательная статистика»

Вспомним и посмотрим еще раз на рисунок 10.Это диаграмма ящика с усами. И сможем отчётливо увидеть выбросы почти во всех данных.

Для дальнейшей более продуктивной работы над данными попробуем убрать эти выбросы используя определенный алгоритм действий.

16

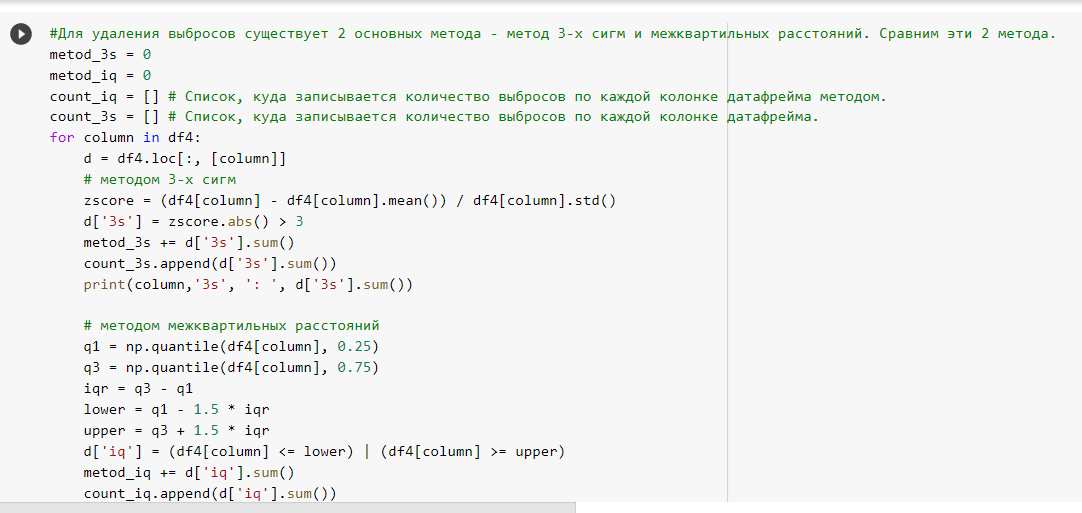
Ниже на рисунке 15, из фрагмента кода видно количество выбросов по каждому столбцу.



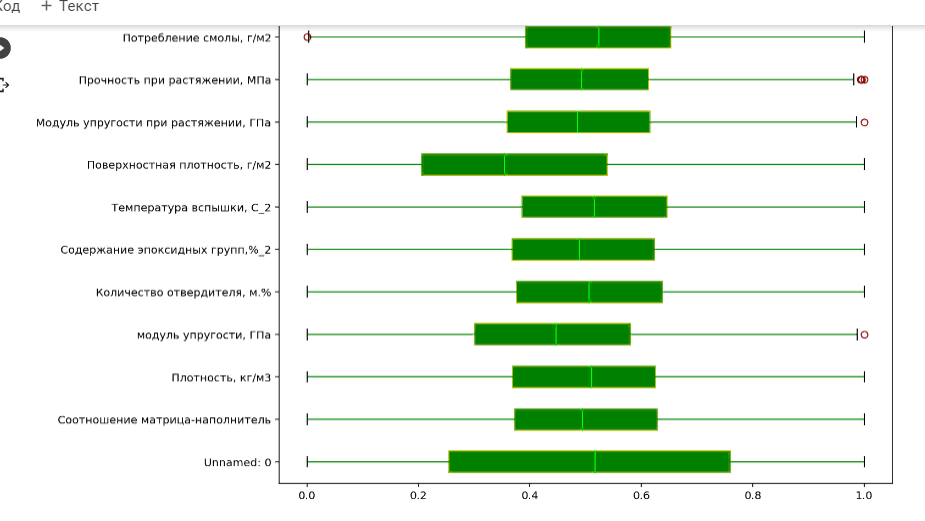
«Рисунок 15. Данные о выбросах»

17

На рисунке 16 представлен фрагмент программного кода для удаления выбросов с помощью двух основных методов таких как, метода трех сигм и межквартальных расстояний.



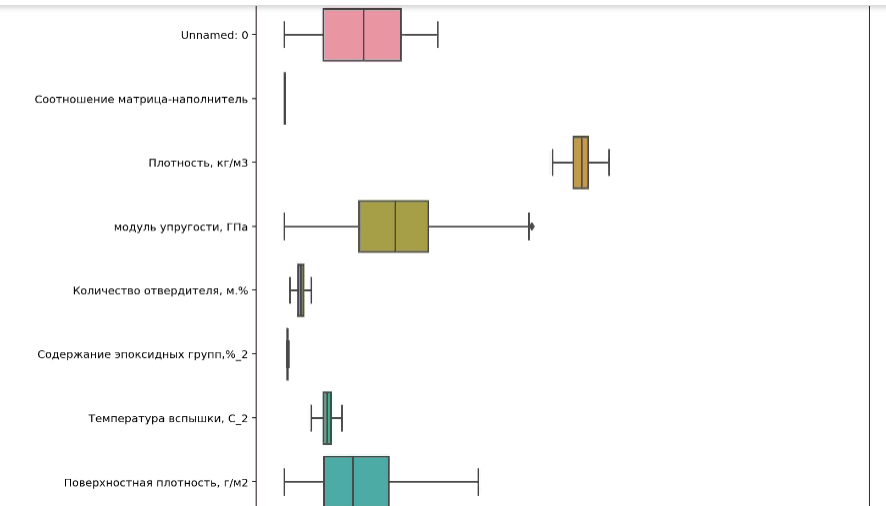
«Рисунок 16. Удаление выбросов»



«Рисунок 17.Диаграмма ящик с усами»

18

Далее на рисунке 17 и 18,после применения методов трех сигм и межквартальных расстояний можно увидеть значительно существенные, визуально видимые изменения практически во всех данных.



«Рисунок 18. Диаграмма ящик с усами»

19

**2.2. Разработка и обучение модели**

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении». Для решения применим все методы, описанные выше. Это Случайный лес, Линейная регрессия, К-ближайших соседей и Дерево решений.

Порядок разработки модели для каждого параметра и для каждого выбранного метода можно разделить на следующие этапы: разделение нормализованных данных на обучающую и тестовую выборки, в соотношении семьдесят на тридцать процентов.

Проверка моделей при стандартных значениях, сравнение с результатами модели, выдающей среднее значение, создание графика, сравнение моделей по метрике МАЕ, поиск сетки гиперпараметров, по которым будет происходить оптимизация модели.

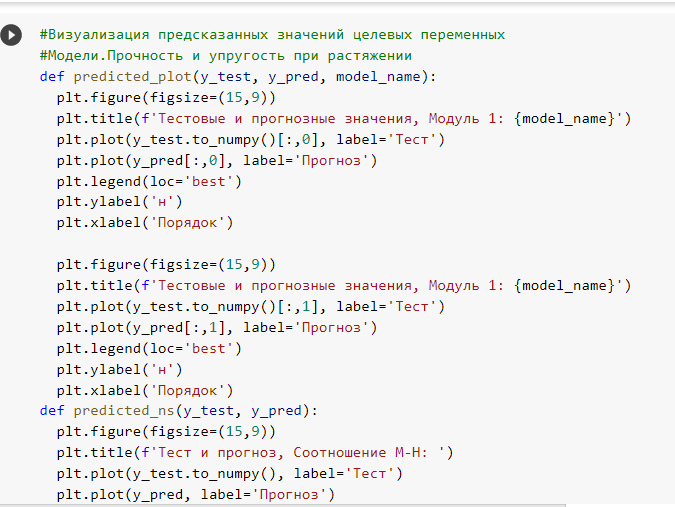
В качестве параметра оценки выбран коэффициент детерминации (R2); оптимизация подбора гиперпараметров модели с помощью выбора по сетке и перекрёстной проверки, подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных, оценка полученных данных, сравнение со стандартными значениями.

В работе, для прогноза модуля упругости при растяжении и прочности были использованы модели такие как: сетчатая регрессия,Grid Search, дерево решений и SVR (регрессия опорных векторов).

20

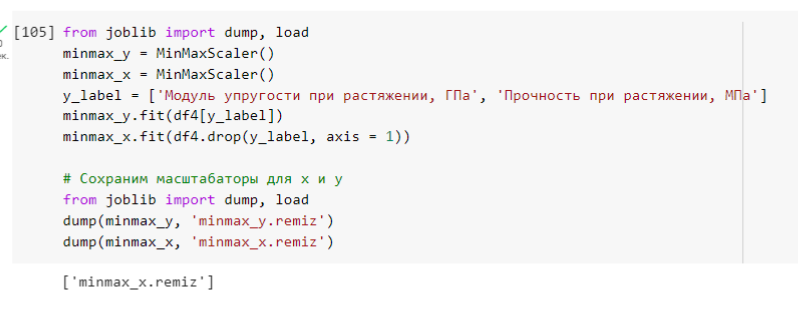
**2.3. Тестирование модели.**

После обучения моделей, была проведена оценка точности этих моделей, на обучающей и тестовых выборках.



«Рисунок 19. Модели»

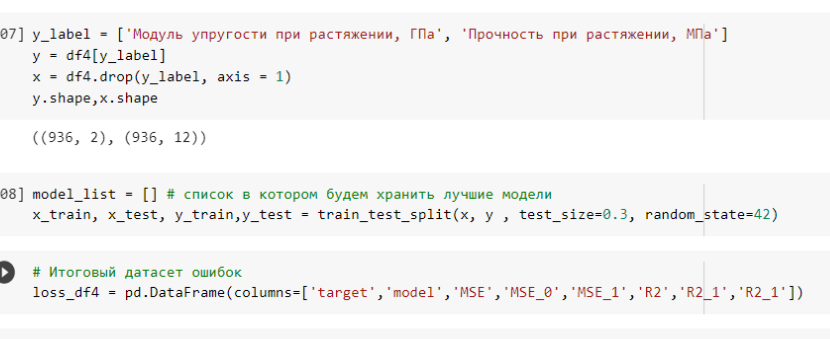
21



«Рисунок 20. Входы и выходы для моделей»

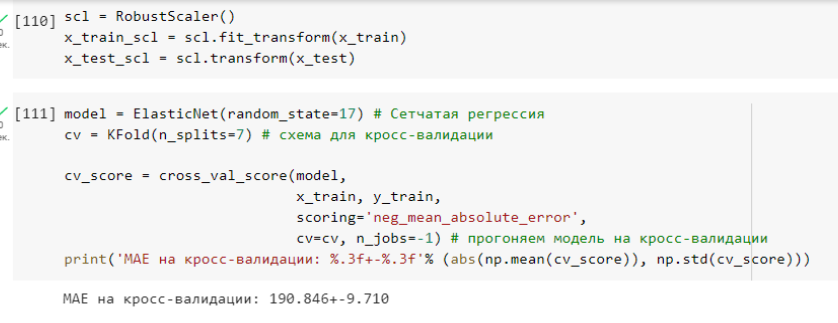


«Рисунок 21. Проверка модели remiz»

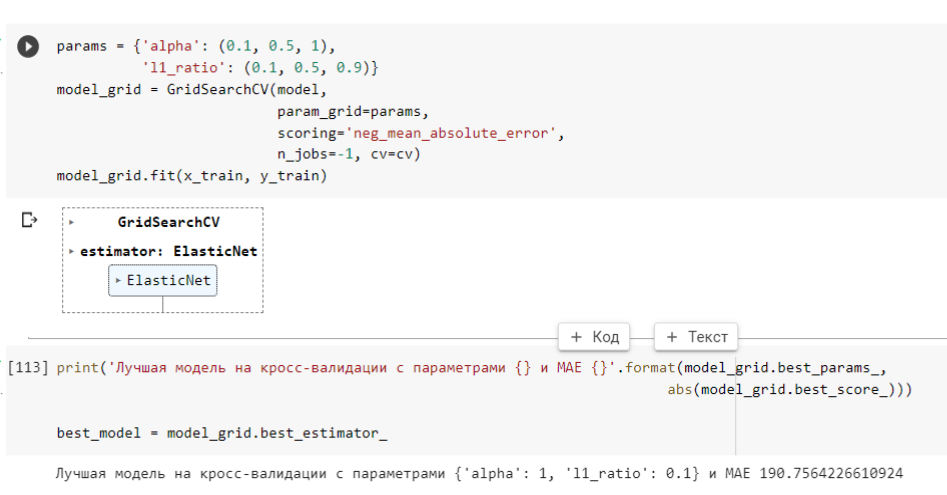


«Рисунок 22. Обучающая и тестовая выборка»

22

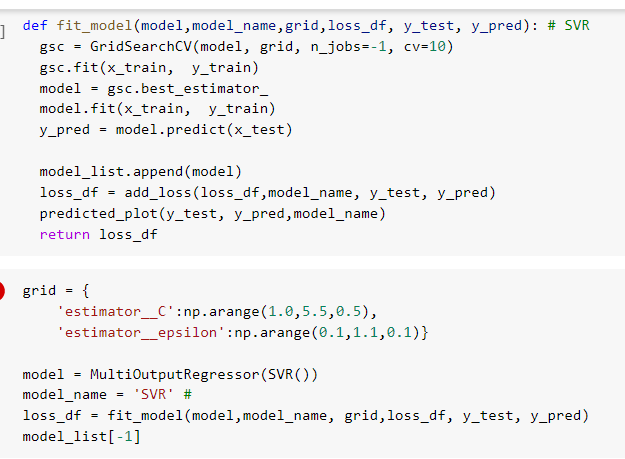


«Рисунок 23. Модель регрессии»



«Рисунок 24.Создание и лучшая модель»

23

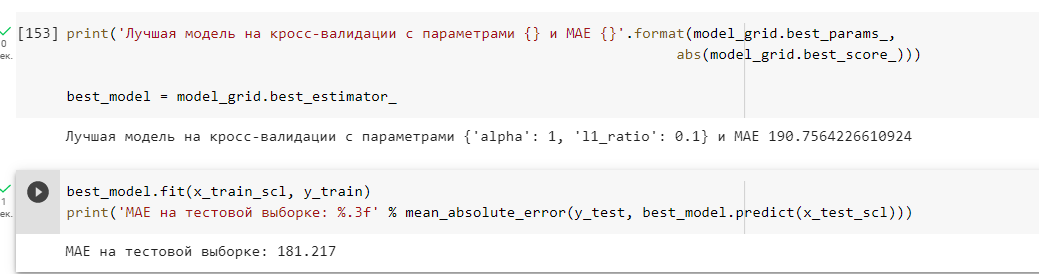


«Рисунок 25. Модель»

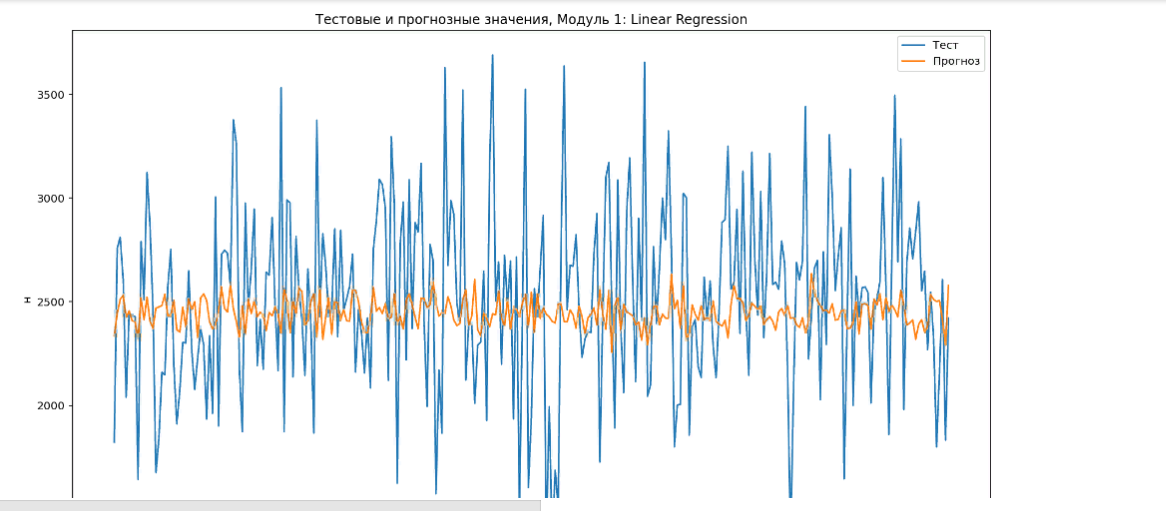


«Рисунок 26.Модель»

24

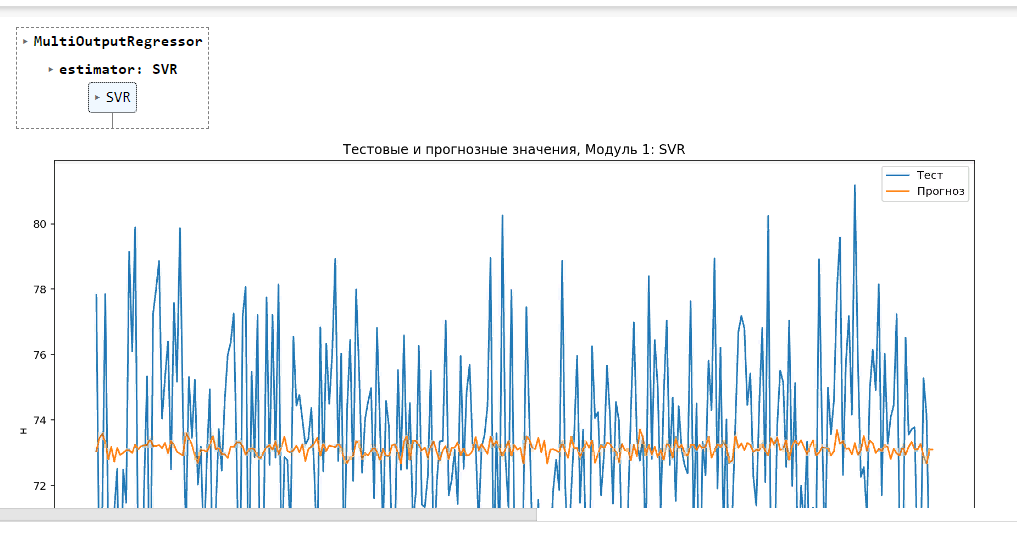


«Рисунок 27. MAE»

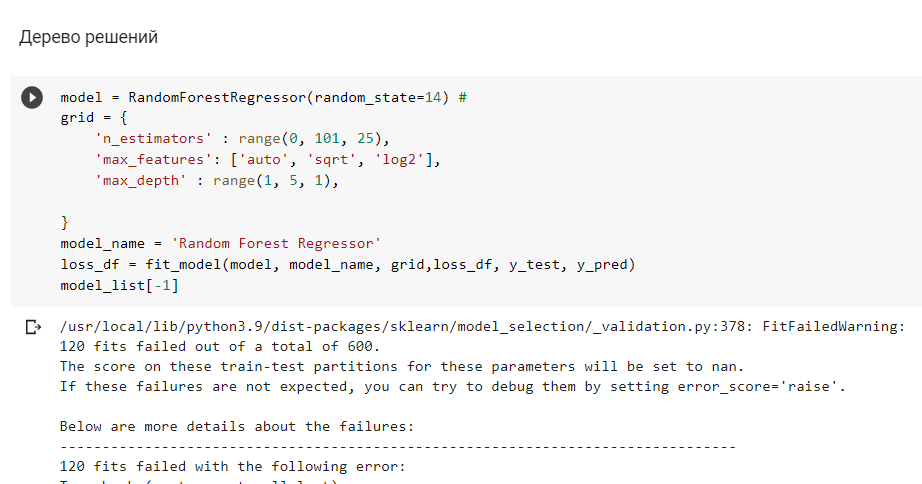


«Рисунок 28»

25

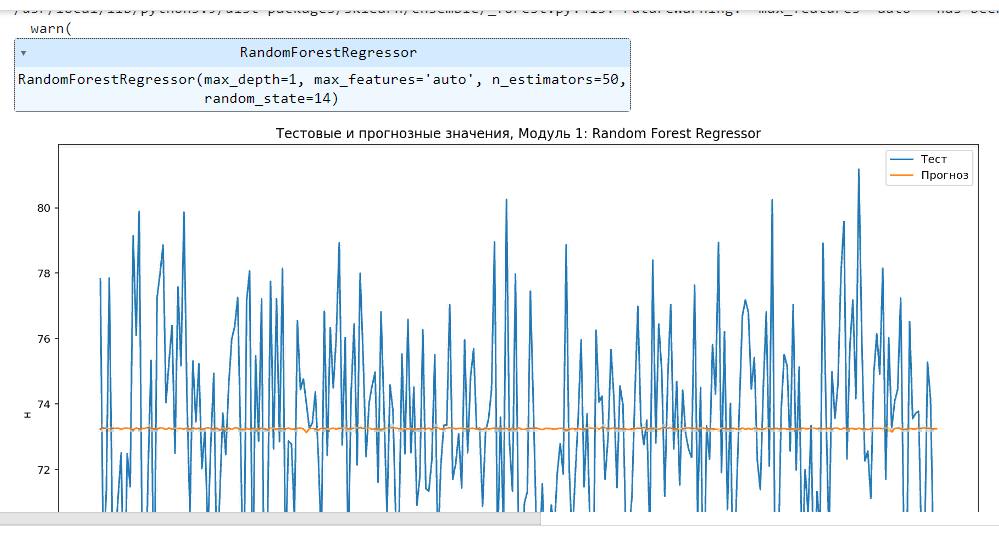


«Рисунок 29»

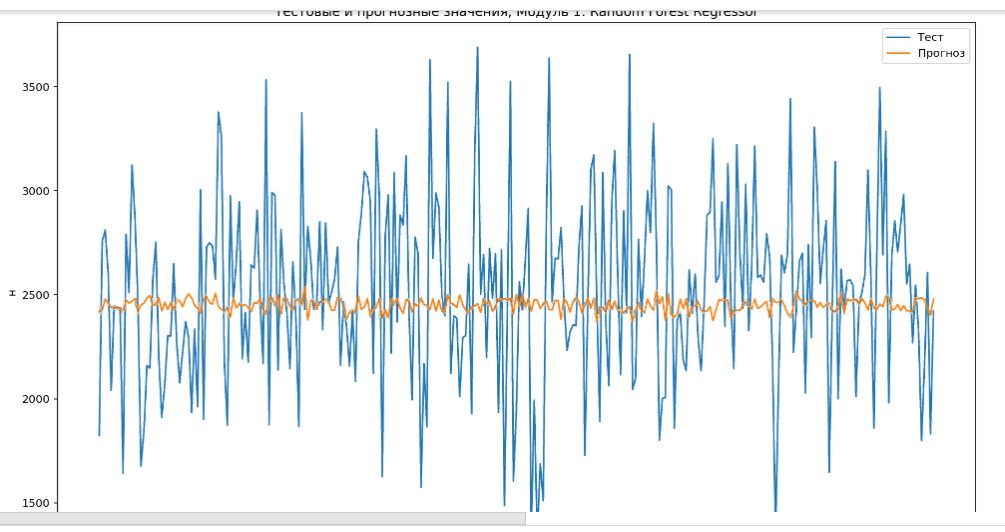


«Рисунок 30.Модель»

26



«Рисунок 31»



«Рисунок 32»

27

Все использованные модели не справились с задачей. Результат неудовлетворительный. Свойства композитных материалов в первую очередь зависят от используемых материалов.

28

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

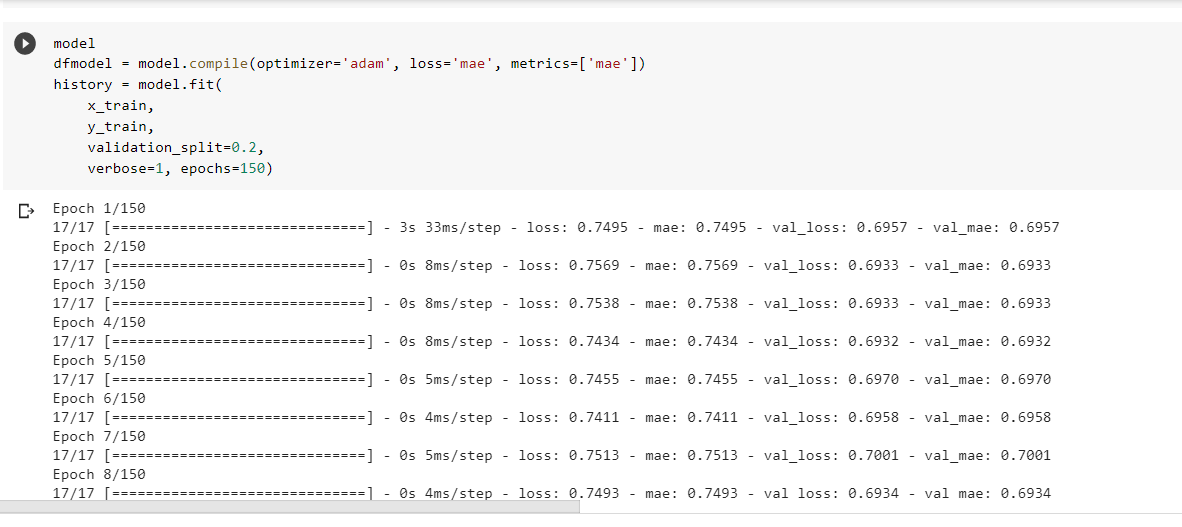
Обучение нейронной сети — это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки.

Начнём стоить нейронную сеть с помощью класса Sequential.

****

«Рисунок 33.Нейронная сеть»

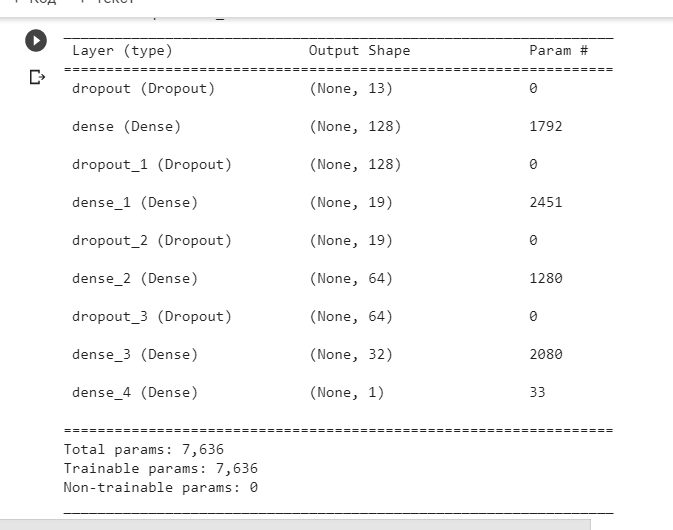
29

****

«Рисунок 34.Нейронная сеть»

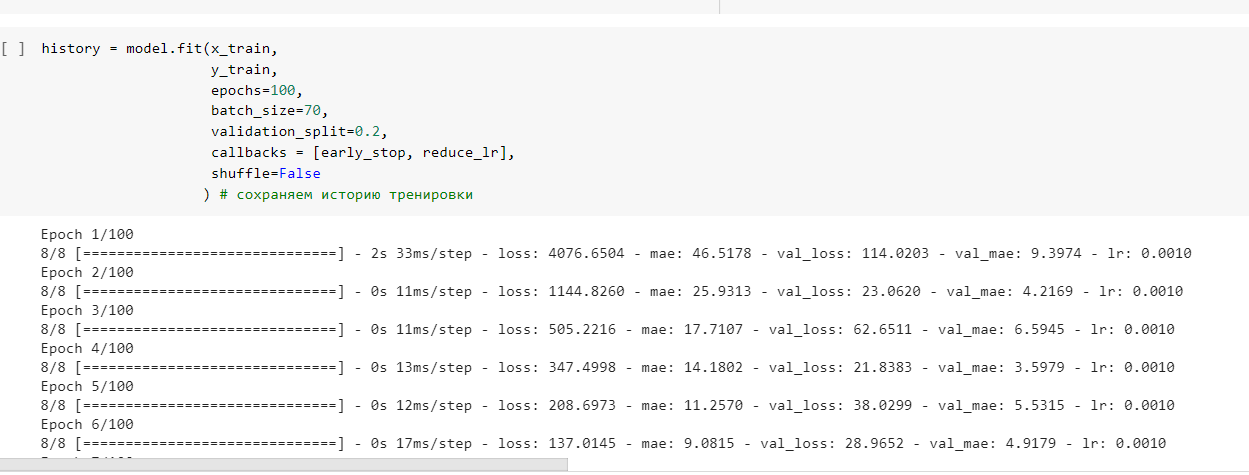
30

Определим параметры, поищем оптимальные параметры, посмотрим на результаты. С помощью KerasClassifier выйдем на наилучшие параметры для нашей нейронной сети и построим окончательную нейросеть.

****

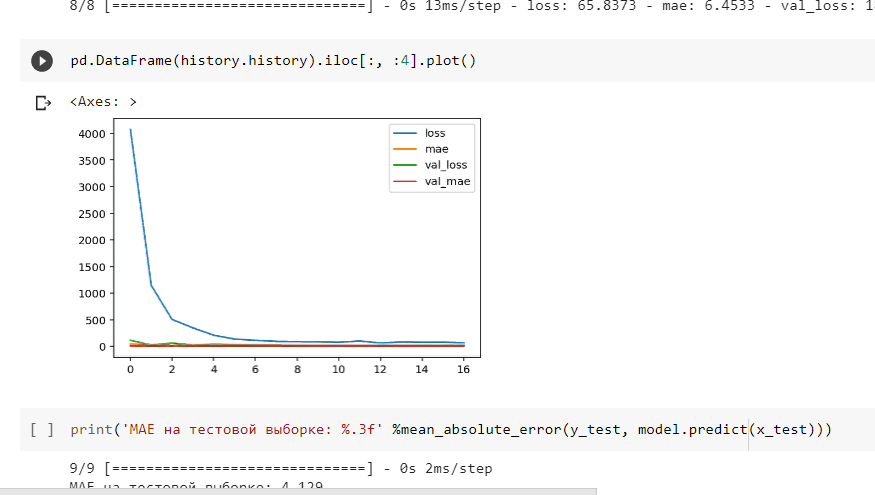
«Рисунок 35»

31



«Рисунок 36.Нейронная сеть»

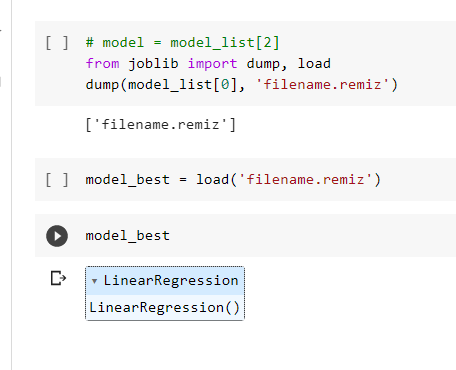
Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.



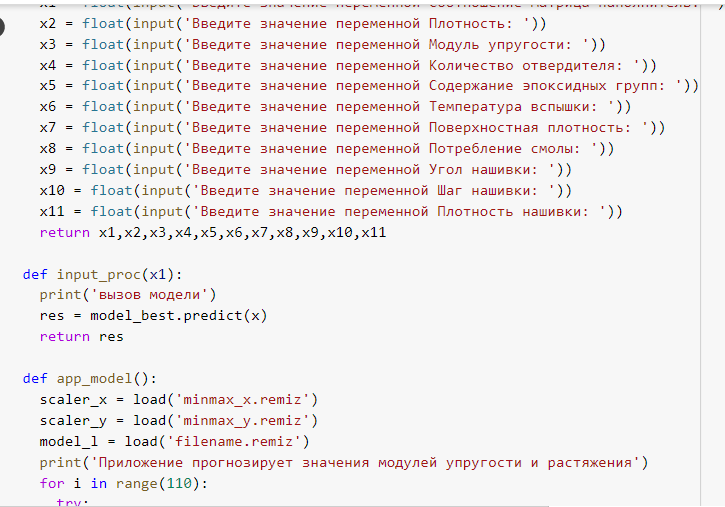
«Рисунок 37»

32

**2.5.Разработка приложения.**

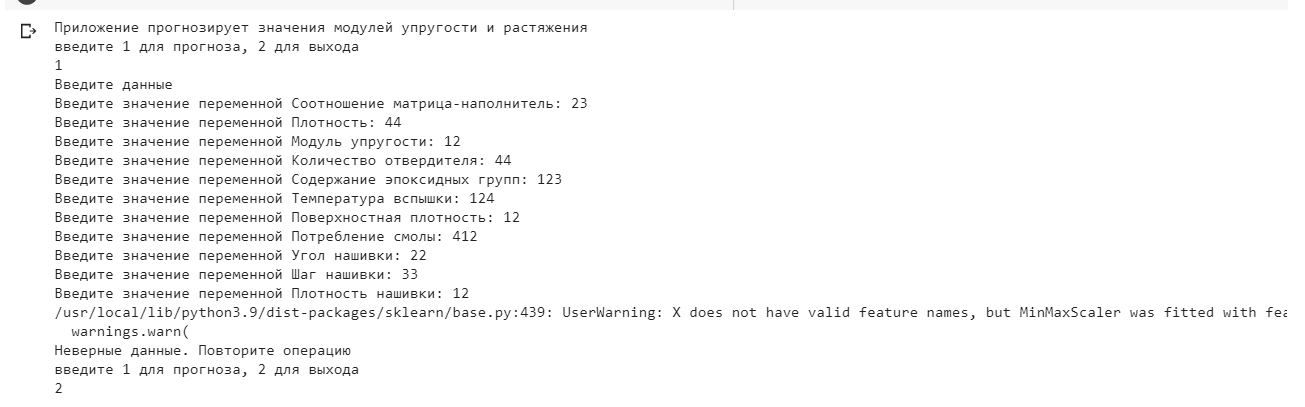
****

«Рисунок 38.Сохранение модели»



«Рисунок 39.Приложение с командной строкой»

33



«Рисунок 40.Работа приложения»

34

**2.6.Создание удаленного репозитория и загрузка результатов работы на**

**него.**

Ссылка на блокнот https://colab.research.google.com/drive/10PplkHbydSBORWsc-dA5sKN3bluJJOqE#scrollTo=D067PY2xbNae

35

**Список используемой литературы и веб ресурсы**.

1. Билл Любанович. Простой Pyton.Современный стиль програмирования.Санкт-Петербург 2016 год. 480 с.
2. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: <https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19>. (дата обращения: 07.06.2022)
3. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.06.2022).
4. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
5. Абросимов Н.А.: Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010
6. Абу-Хасан Махмуд, Масленникова Л. Л.: Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год
7. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.
8. Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
9. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
10. Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.(дата обращения: 08.06.2022).
11. Документация по библиотеке matplotlib: – Режим доступа: <https://matplotlib.org/stable/users/index.html>. (дата обращения: 10.06.2022)
12. Документация по библиотеке numpy: – Режим доступа: <https://numpy.org/doc/1.22/user/index.html#user>. (дата обращения: 03.06.2022).
13. Документация по библиотеке pandas: – Режим доступа: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>. (дата обращения: 04.06.2022).
14. Документация по библиотеке scikit-learn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>. (дата обращения: 05.06.2022).
15. Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>. (дата обращения: 06.06.2022).
16. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview> (дата обращения: 10.06.2022).
17. Документация по языку программирования python: – Режим доступа:  <https://docs.python.org/3.8/index.html>. (дата обращения: 02.06.2022).
18. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
19. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: <https://habr.com/ru/post/428503/> (дата обращения 07.06.2022)
20. Ларин А. А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.
21. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
22. Миронов А.А. Машинное обучение часть I ст.9 – Режим доступа: <http://is.ifmo.ru/verification/machine-learning-mironov.pdf>. (дата обращения 08.06.2022)
23. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.
24. Реутов Ю.А.: Прогнозирование свойств полимерных композиционных материалов и оценка надёжности изделий из них, Диссертация на соискание учёной степени кандидата физико-математических наук, Томск 2016.
25. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.
26. Руководство по быстрому старту в flask: – Режим доступа: <https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html>. (дата обращения: 09.06.2022)
27. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
28. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.
29. Справочник по композиционным материалам: в 2 - х кн. Кн. 2 / Под ред. Дж. Любина; Пер. с англ. Ф. Б. Геллера, M. М. Гельмонта; Под ред. Б. Э. Геллера - М.: Машиностроение, 1988. - 488 с. : ил;
30. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с.: ил.
31. Чун-Те Чен и Грейс Х. Гу. Машинное обучение для композитных материалов (март 2019г.) – Режим доступа: [https://www.cambridge.org/core/journals/mrs-communications/article/machine- learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15](https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15). (дата обращения 02.06.2022)