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**

«DataScience»

Клещев Сергей Владимирович

Слушатель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (ФИО)

Москва, 2022

**Содержание**

Введение1

1. Аналитическая часть2

1.1 Постановка задачи2

* 1. Описание используемых методов11
  2. Разведочный анализ17

1. Практическая часть21
   1. Предобработка данных21

2.2Разработка и обучение модели27

* 1. Тестирование модели29

2.4Нейронная сеть32

2.5 Создание удаленного репозитория и загрузка результатов работы на него34

Вывод35

Список литературы36

**Введение**

**Наука о данных** англ. *data science*; иногда *даталогия* *datalogy*) раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интелекта для работы с данными, а также методы проектирования и разработки баз данных.

Рассматривается как академическая дисциплина, а с начала 2010-х годов, во многом благодаря популяризации концепции «больших данных» — и как практическая межотраслевая сфера деятельности, притом специализация исследователя данных (англ. *data scientist* — «учёного по данным») с начала 2010-х годов считается одной из самых привлекательных, высокооплачиваемых и перспективных профессий.

© Wikipedia

**Специалист в области Data Science строит на основе данных модели, которые помогают принимать решения в науке, бизнесе и повседневной жизни.** Он может работать с неструктурированными массивами информации в разных сферах: от выявления элементарных частиц в экспериментах на БАК, анализа метеорологических факторов, анализа данных о перемещениях автотранспорта до исследования финансовых операций, поисковых запросов, поведения пользователей в Интернете.

В результате получаются модели, которые прогнозируют погоду, загруженность дорог, спрос на товары, находят снимки, где могут оказаться следы нужных элементарных частиц, выдают решения о предоставлении кредита, могут рекомендовать товар, книгу, фильм, музыку.

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

Целью задачи является прогнозирование конечных свойств композиционных материалов (базальтопластик) с применением методов DataScience в среде программирования Python.

Базальтопластик представляет собой конструкцию композиционного порядка, состоящую из стеклянного наполнителя, а также из полимерного синтетического связующего. Наполнителем, который выполняет скрепляющую и армирующую функции, служат стеклянные волокна, сделанные в виде жгутов и нитей. Это стекловолокно обрабатывается особым связующим материалом, способным обеспечить высокую прочность на сдвиг и увеличить сопротивление разрушающему воздействию агрессивных химических веществ.

В наши дни композитная арматура находит все больше и больше места в строительстве, она используется во многих областях промышленности и производства. Причина подобной популярности материала кроется в его очень высокой прочности и стойкости к коррозии, низкой теплопроводности и высокой морозостойкости.  Конструкции из стеклопластика и базальтопластика не электропроводны, что имеет большую важность для исключения в материалах блуждающих токов.

Для исследований даны датасеты «X\_bp.xlsx» и «X\_nup.xlsx». В файле «X\_bp.xlsx» содержится таблица с характеристиками материалов по 10 свойствам из 1023 значений, а в файле «X\_nup.xlsx» содержится таблица, описывающая характеристики материалов нашивки по 3 свойствам из 1040 значений.

Предоставленные данные:

**Входные переменные:**

Соотношение матрица-наполнитель

Плотность, кг/м3

Модуль упругости, Гпа

Количество отвердителя, м

Содержание эпоксидных групп,%\_2

Температура вспышки, С\_2

Поверхностная плотность, г/м2

Модуль упругости при растяжении, Гпа

Прочность при растяжении, Мпа

Потребление смолы, г/м2

Угол нашивки, град

Шаг нашивки

Плотность

**Выходные переменные (исключаются в момент решения задачи из входных):**

**Задача регрессии 1:**

Модуль упругости при растяжении, Гпа

**Задача регрессии 2:**

Прочность при растяжении, Мпа

**Разработка рекомендательной системы:**

Соотношение матрица-наполнитель

**Первый шаг в обработке данных:**

Объединение датасетов по индексу с применением метода INNER

При анализе найдены выбросы в каждом из признаков, кроме «Угол нашивки, град».

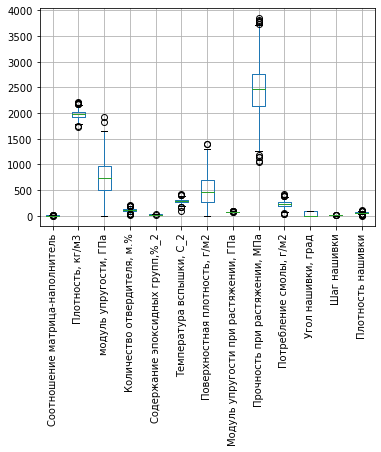
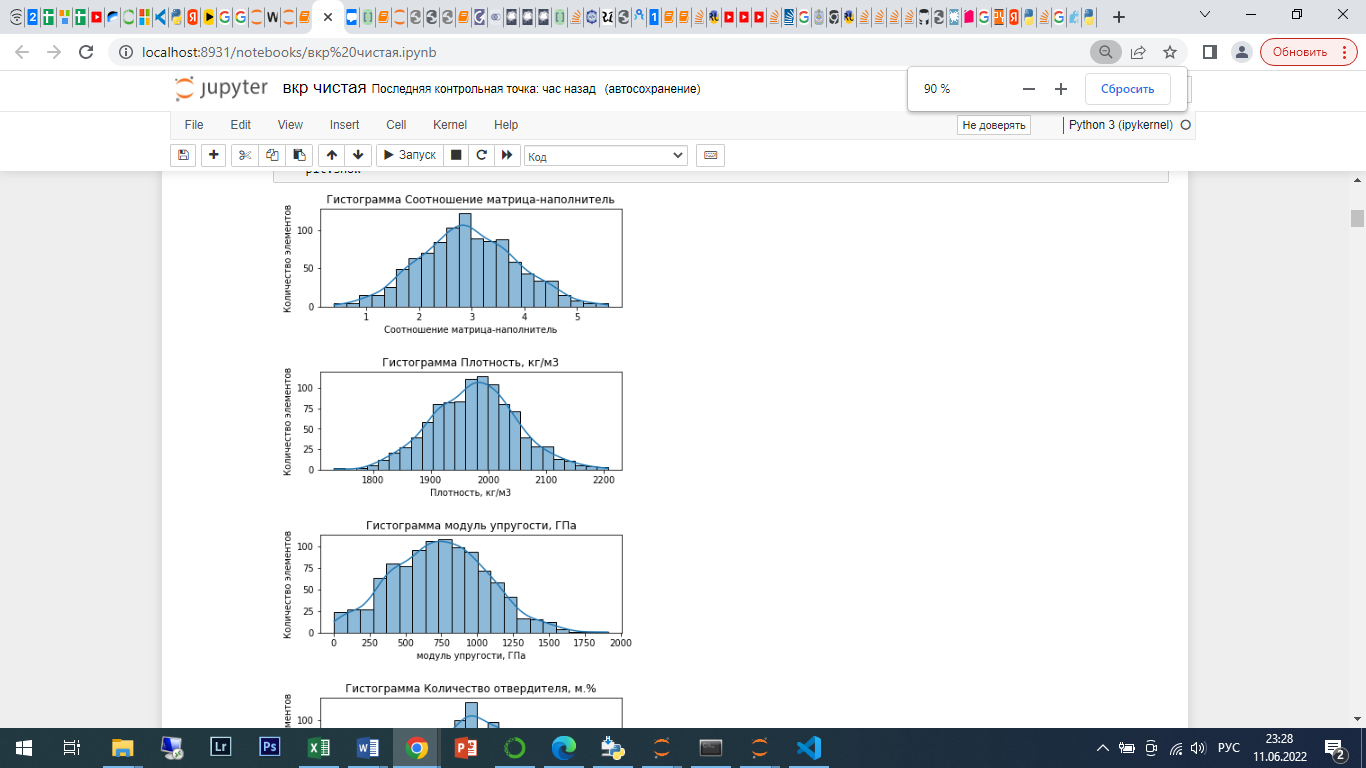
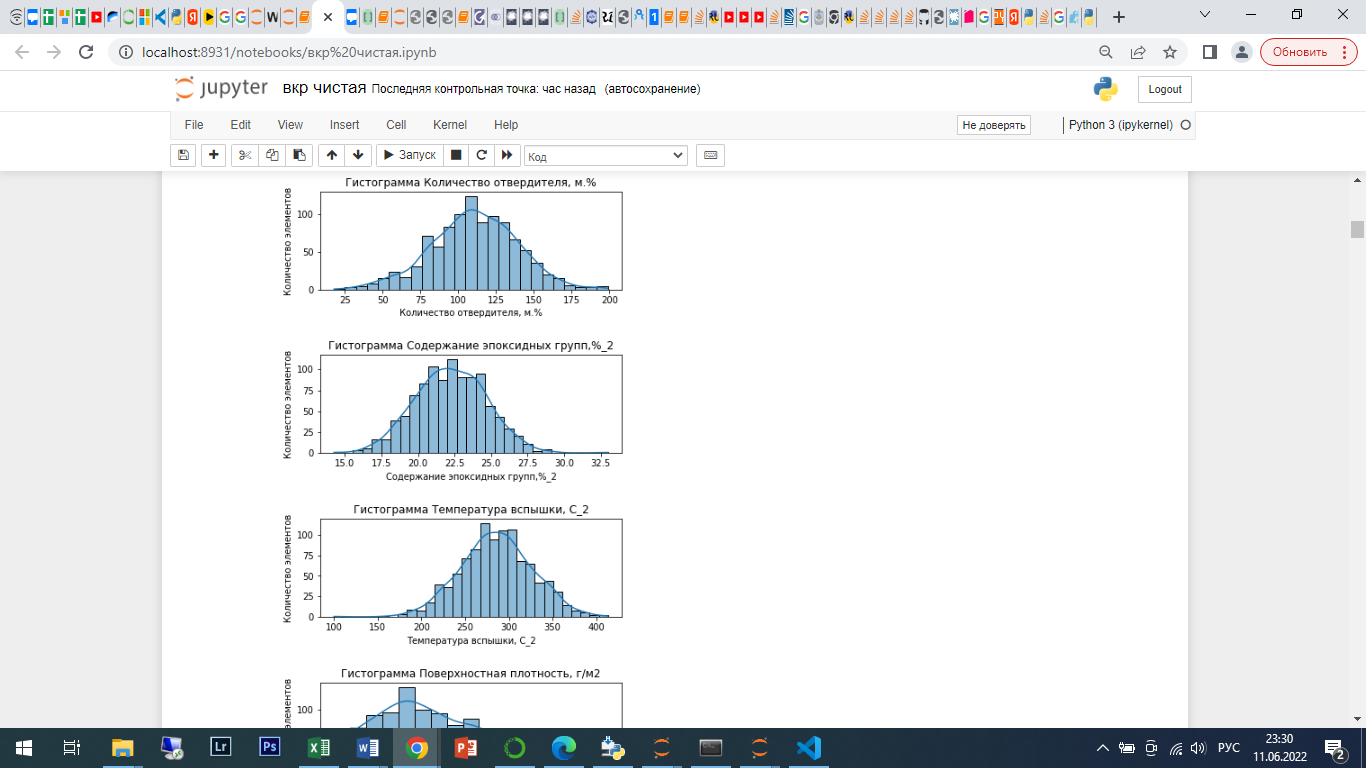
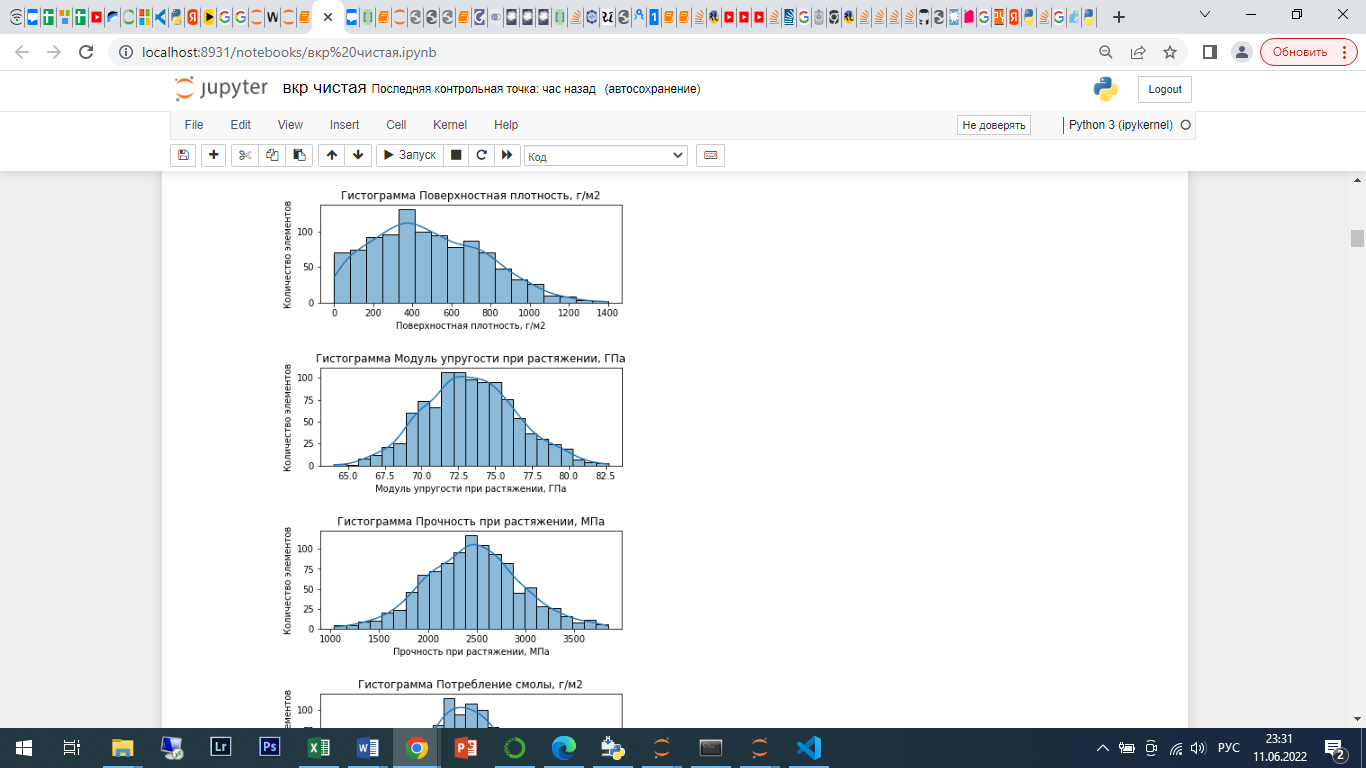


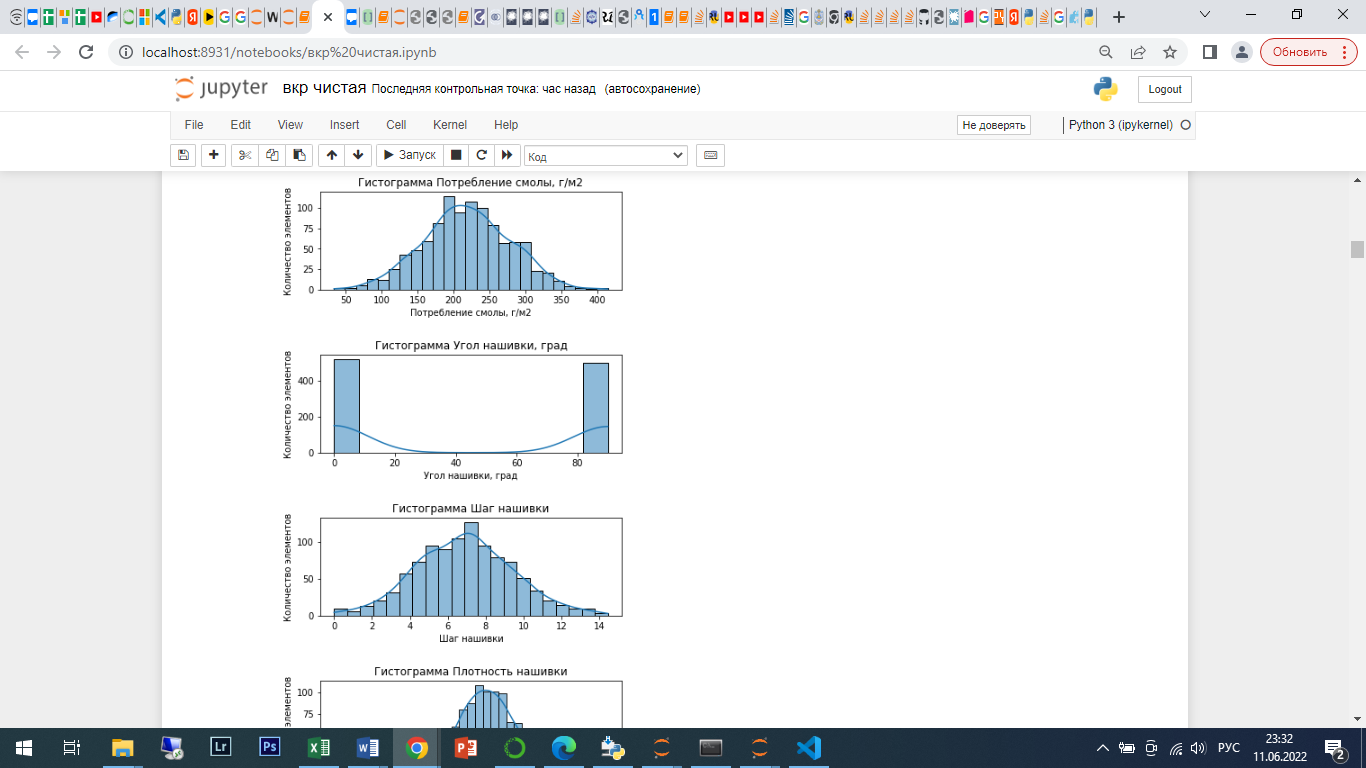
Рисунок 1 – Ящик с усами

Гистограммы распределения значений до момента проведения каких- либо операций представлены на рисунке 2.









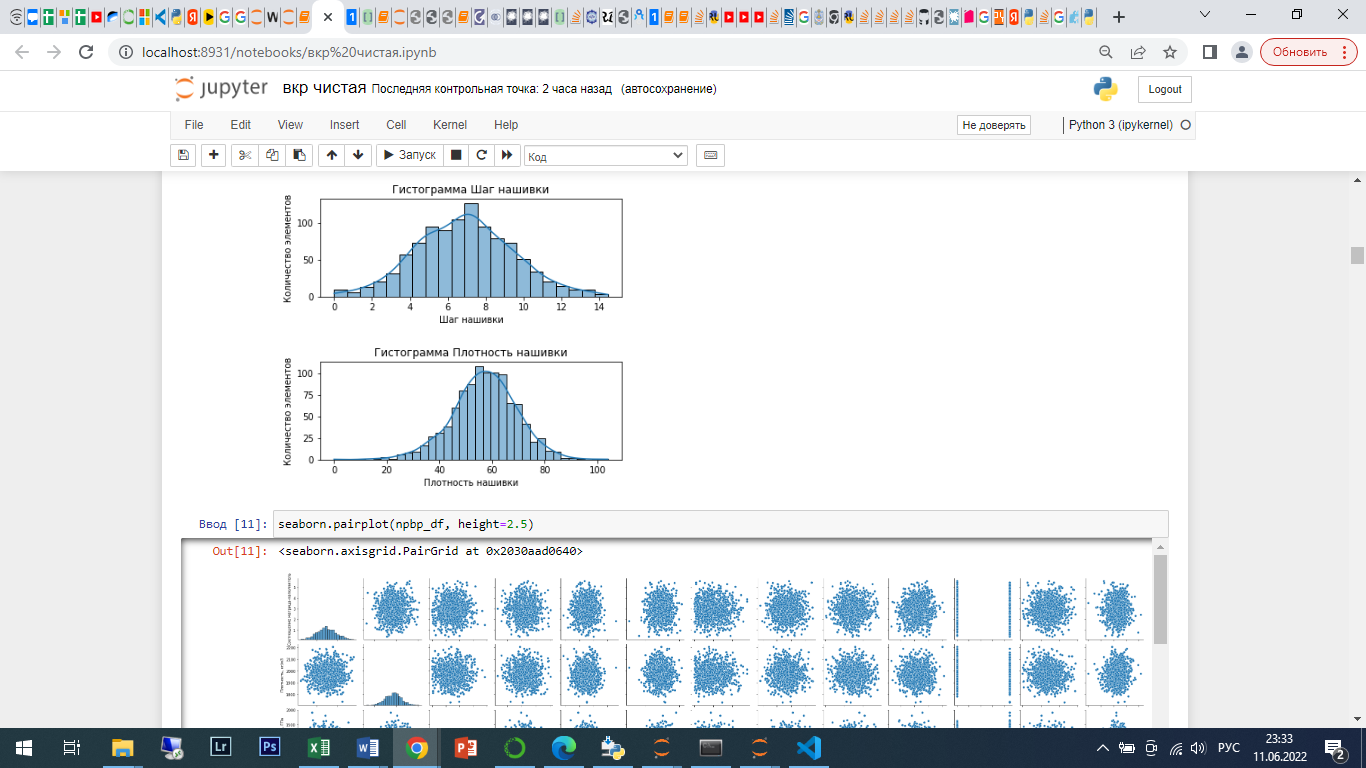


Рисунок 2 – Гистограммы распределения изначальных данных

На данных графиках мы видим большое количество выбросов (наибольшее в параметре «Плотность нашивки»), где будет необходимо выбрать метод, который поможет нам от них избавиться.

Соотношение матрица-наполнитель 0

Плотность, кг/м3 52

модуль упругости, ГПа 35

Количество отвердителя, м.% 49

Содержание эпоксидных групп,%\_2 6

Температура вспышки, С\_2 43

Поверхностная плотность, г/м2 16

Модуль упругости при растяжении, ГПа 12

Прочность при растяжении, МПа 61

Потребление смолы, г/м2 46

Угол нашивки, град 0

Шаг нашивки 8

Плотность нашивки 55

* 1. **Описание используемых методов**

а) Удаление выбросов по методу квартилей

A квартиль - это тип квантиля, который делит количество точек данных на четыре более или менее равных части или четверти. Первый квартиль (Q 1 ) определяется как среднее число между наименьшим числом и медианой набора данных. Он также известен как нижний квартиль или 25-й эмпирический квартиль и отмечает, где 25% данных находятся ниже или слева от него (если данные упорядочены на временной шкале от наименьшего к наибольшему). Второй квартиль (Q 2 ) - это медиана набора данных, и 50% данных находится ниже этой точки. Третий квартиль (Q 3 ) - это среднее значение между медианой и наивысшим значением набора данных. Он также известен как верхний квартиль или 75-й эмпирический квартиль, и 75% данных находятся ниже этой точки. В связи с тем, что для вычисления квартилей необходимо упорядочить данные от наименьшего к наибольшему, квартили представляют собой форму статистики порядка.

Наряду с минимальным и максимальным значением данных, которые также являются квартилями, три Описанные выше квартили предоставляют сводку из пяти цифр данных. Эта сводка важна для статистики, поскольку она предоставляет информацию как о центре, так и о разбросе данных. Знание нижнего и верхнего квартилей дает информацию о том, насколько велик разброс и смещен ли набор данных в одну сторону. Поскольку квартили делят количество точек данных поровну, диапазон между квартилями неодинаков (т. Е. Q 3-Q2≠ Q 2-Q1). В то время как максимум и минимум также показывают разброс данных, верхний и нижний квартили могут предоставить более подробную информацию о расположении конкретных точек данных, наличии выбросов в данных и разнице в разбросе между средними 50% данных. данные и внешние точки данных.

б) Линейная регрессия (Linearregression)

Модель линейной регрессии является часто используемой и наиболее изученной в [эконометрике](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0). А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями. С эконометрической точки зрения более важное значение имеет [линейность по параметрам](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%BF%D0%BE_%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B0%D0%BC), чем линейность по факторам модели.

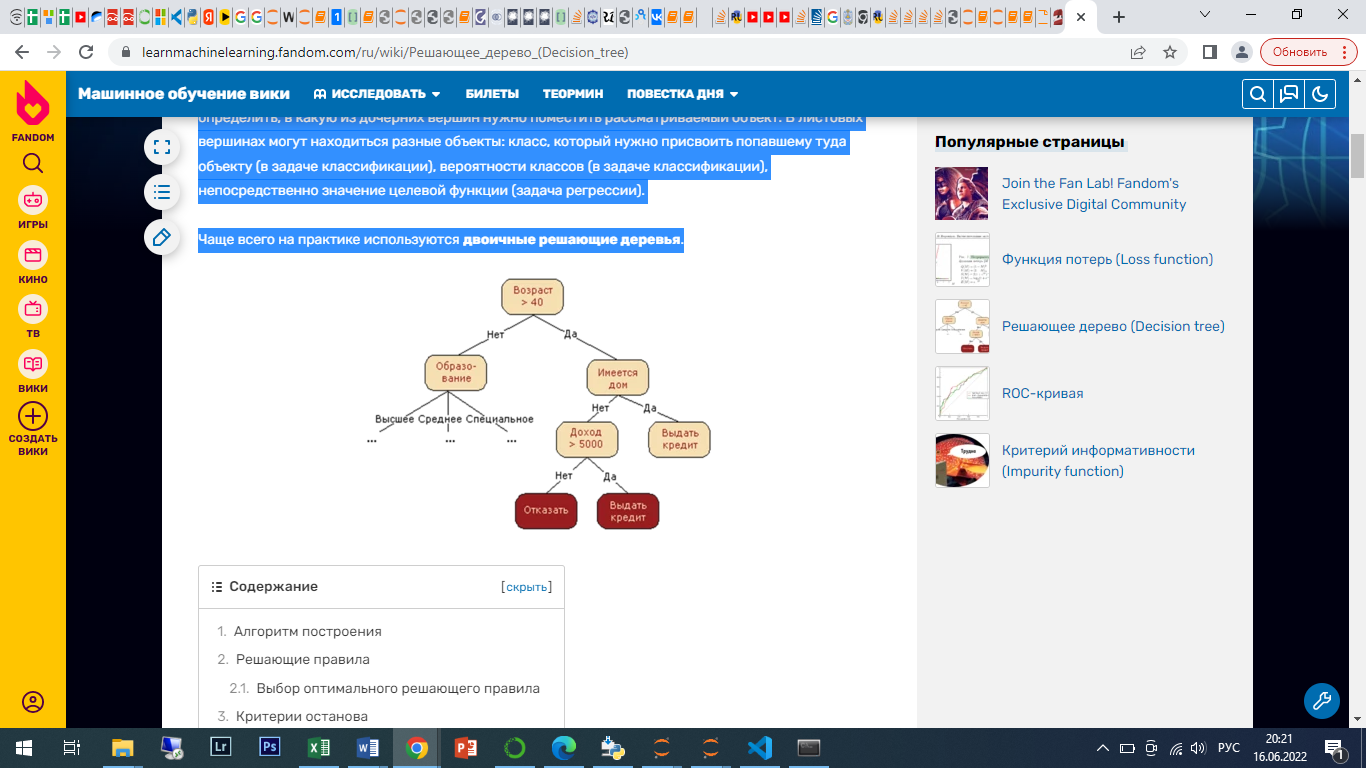
в) Метод k-ближайших соседей (K-nearestneighbor)

**Метод k**-**ближайших соседей** (**k**-nearest neighbors) – это простой алгоритм машинного обучения с учителем, который можно использовать для решения задач классификации и регрессии. Он прост в реализации и понимании, но имеет существенный недостаток – значительное замедление работы, когда объем данных растет.

г) Ансамбли решателей. Решающие деревья (Decisiontree)

Решающее дерево (Decisiontree) — решение задачи [обучения с учителем](https://learnmachinelearning.fandom.com/ru/wiki/%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC_(Supervised_learning)), основанный на том, как решает задачи прогнозирования человек. В общем случае — это k-ичное дерево с решающими правилами в нелистовых вершинах (узлах) и некотором заключении о целевой функции в листовых вершинах (прогнозом). Решающее правило — некоторая функция от объекта, позволяющее определить, в какую из дочерних вершин нужно поместить рассматриваемый объект. В листовых вершинах могут находиться разные объекты: класс, который нужно присвоить попавшему туда объекту (в задаче классификации), вероятности классов (в задаче классификации), непосредственно значение целевой функции (задача регрессии).

Чаще всего на практике используются двоичные решающие деревья.



Преимущество метода — сокращение временных затрат на обучение, также еще из основных преимуществ решающих деревьев является возможность работы с пропущенными значениями. Деревья решений создаются по понятным правилам, они просты в применении и интерпретации. Способны работать с большим количеством объектов и признаков, имеют небольшое количество параметров, легко регуляризуются. Недостатки: обобщающая способность решающих деревьев невысока, нет онлайновости - для новых наблюдений требуется полная перестройка всего дерева, они могут восстанавливать только линейные зависимости между целевой переменной и признаками

Анализ качества получившихся моделей с помощью коэффициент детерминации

Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным.

Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50 % (в этом случае коэффициент множественной корреляции превышает по модулю 70 %). Модели с коэффициентом детерминации выше 80 % можно признать достаточно хорошими (коэффициент корреляции превышает 90 %). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

**Использование нейронной сети для прогнозирования**

Персептрон - математическая модель нейронной сети, состоящего из одного нейрона, который выполняет две последовательные операции: вычисляет сумму входных сигналов с учетом их весов (проводимости или сопротивления) связи и применяет активационную функцию к общей сумме воздействия входных сигналов. В качестве активационной функции может использоваться любая дифференцируемая функция. Выбор активационной функции ложиться на плечи инженера, и обычно этот выбор основан или на уже имеющемся опыте решения похожих задач, ну или просто методом подбора.

Процесс обучения состоит из несколько шагов.

ШАГ 1 — Прямое распространение ошибки (feedforwardprocess)  
На данном шаге вычисляется сумма входных сигналов с учетом веса каждой связи и применяется активационная функция.

ШАГ 2 — Расчет функции ошибки  
Функция ошибка – это метрика, отражающая расхождение между ожидаемыми и полученными выходными данными. Обычно используют следующие функции ошибки: среднеквадратичная ошибка (MeanSquaredError, MSE) – данная функция ошибки особенно чувствительна к выбросам в тренировочном наборе, так как используется квадрат от разности фактического и ожидаемого значений (выброс — значение, которое сильно удалено от других значений в наборе данных, которые могут иногда появляться в следствии ошибок данных, таких как смешивание данных с разными единицами измерения или плохие показания датчиков)

ШАГ 3 — Обратное распространение ошибки (backpropagation)  
Цель обучения нейронной сети проста – это минимизация функции ошибки: Одним способом найти минимум функции – это на каждом очередном шаге обучения модифицировать веса соединений в направлении противоположным вектору-градиенту – метод градиентного спуска.

* 1. **Разведочный анализ данных**

а) Построение гистограмм распределения

В описательной статистике гистограмма распределения — наглядное представление функции плотности вероятности некоторой случайной величины, построенное по выборке. Иногда её называют частотным распределением, так как гистограмма показывает частоту появления измеренных значений параметров объекта.

Высота каждого столбца указывает на частоту появления значений параметров в выбранном диапазоне, а количество столбцов – на число выбранных диапазонов.

Важное преимущество гистограммы заключается в том, что она позволяет наглядно представить тенденции изменения измеряемых параметров качества объекта и зрительно оценить закон их распределения. Кроме того, гистограмма дает возможность быстро определить центр, разброс и форму распределения случайной величины. Строится гистограмма, как правило, для интервального изменения значений измеряемого параметра.

б) Представление выбросов. «Ящик с усами»

Ящик с усами, диаграмма размаха ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) box-and-whiskers diagram or plot, box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное [распределение вероятностей](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9).

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний [квартили](https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D0%BB%D1%8C), минимальное и максимальное значение выборки и [выбросы](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)). Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

График «ящик с усами», или «ящичковая диаграмма», был разработан [Джоном Тьюки](https://ru.wikipedia.org/wiki/%D0%A2%D1%8C%D1%8E%D0%BA%D0%B8,_%D0%94%D0%B6%D0%BE%D0%BD) в 1970-х годах. По сути, ящик с усами — это быстрый способ изучения одного или нескольких наборов данных в графическом виде. Этот график может показаться более примитивным, чем, например, [гистограммы](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D1%81%D1%82%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0), но он имеет некоторые преимущества. Он занимает меньше места и поэтому особенно полезен для сравнения распределений между несколькими группами или наборами данных. Кроме того, ящик с усами в своей первоначальной форме прост для построения.

в) Тепловая карта корреляции (Heatmapcorr)

Корреляция - важнейший фактор, лежащий в основе анализа данных. Он сообщает, как переменные в наборе данных связаны друг с другом и как они перемещаются относительно друг друга. Значение корреляции колеблется от -1 до +1. 0 Корреляция означает, что две переменные не зависят друг от друга. Положительная корреляция указывает на то, что переменные движутся в одном направлении, а отрицательная корреляция указывает на противоположное.

Тепловая карта используется для графического представления значений матрицы с разными цветовыми оттенками для разных значений. Он очень четко визуализирует общую матрицу.

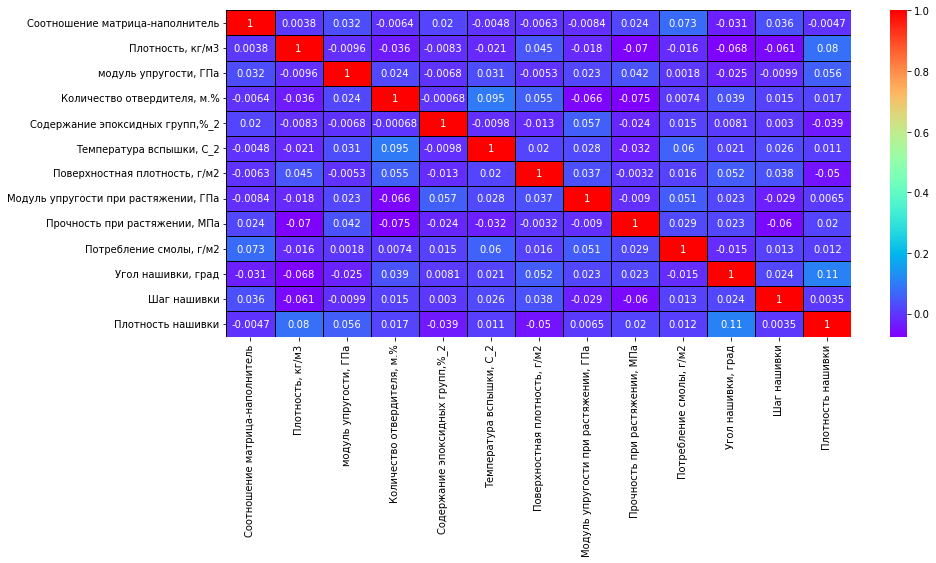


Рисунок 3 – Heatmapcorr данных

г) Pairplot

График pairplot показывает отношения между всеми парами переменных.

Он основан на классе PairGrid в seaborn и возвращает объект его типа. Он также отображает все числовые столбцы данных на обеих осях, отображая матрицу различных графиков. Он дает окончательную фигуру с графиком.

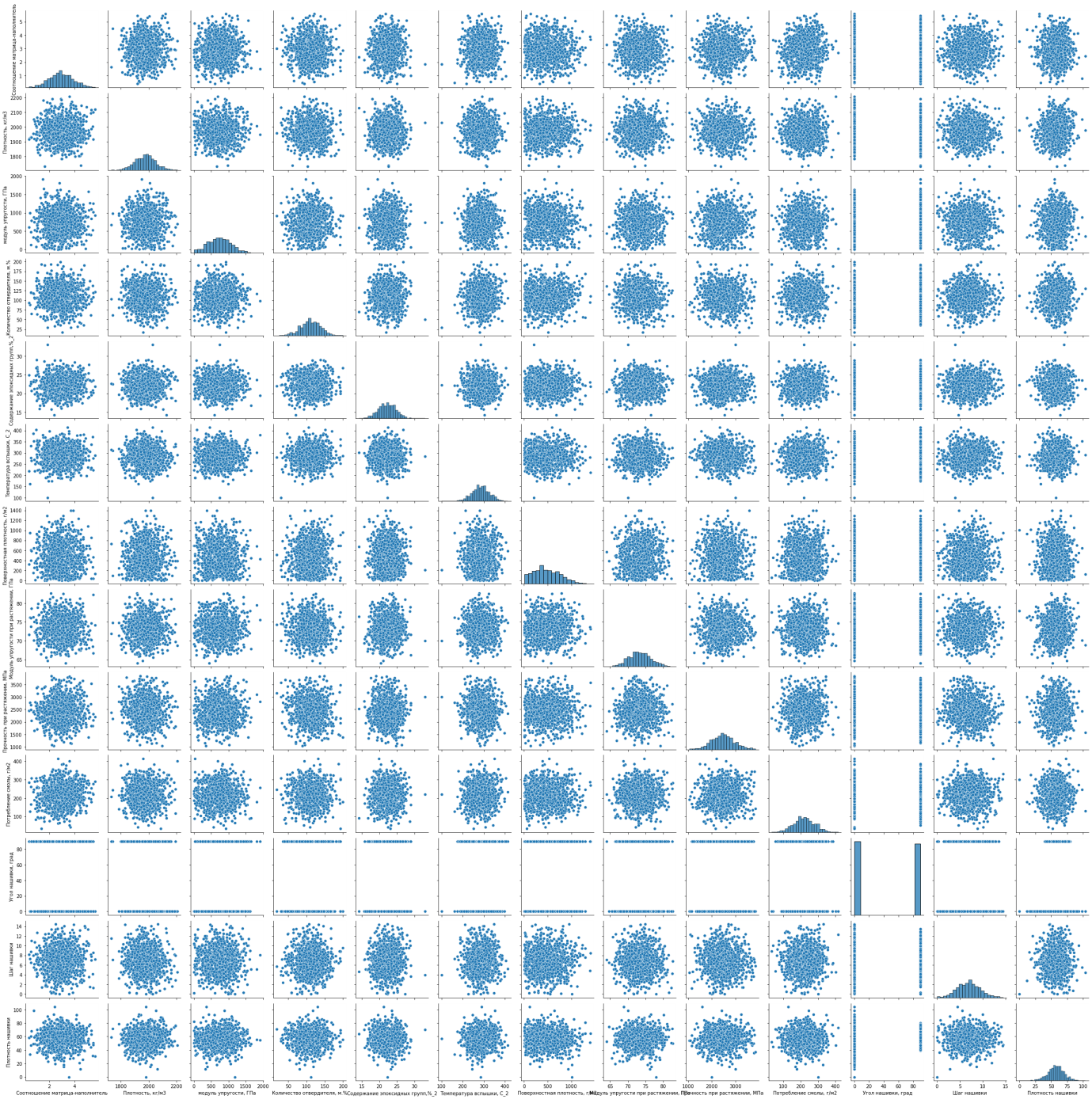
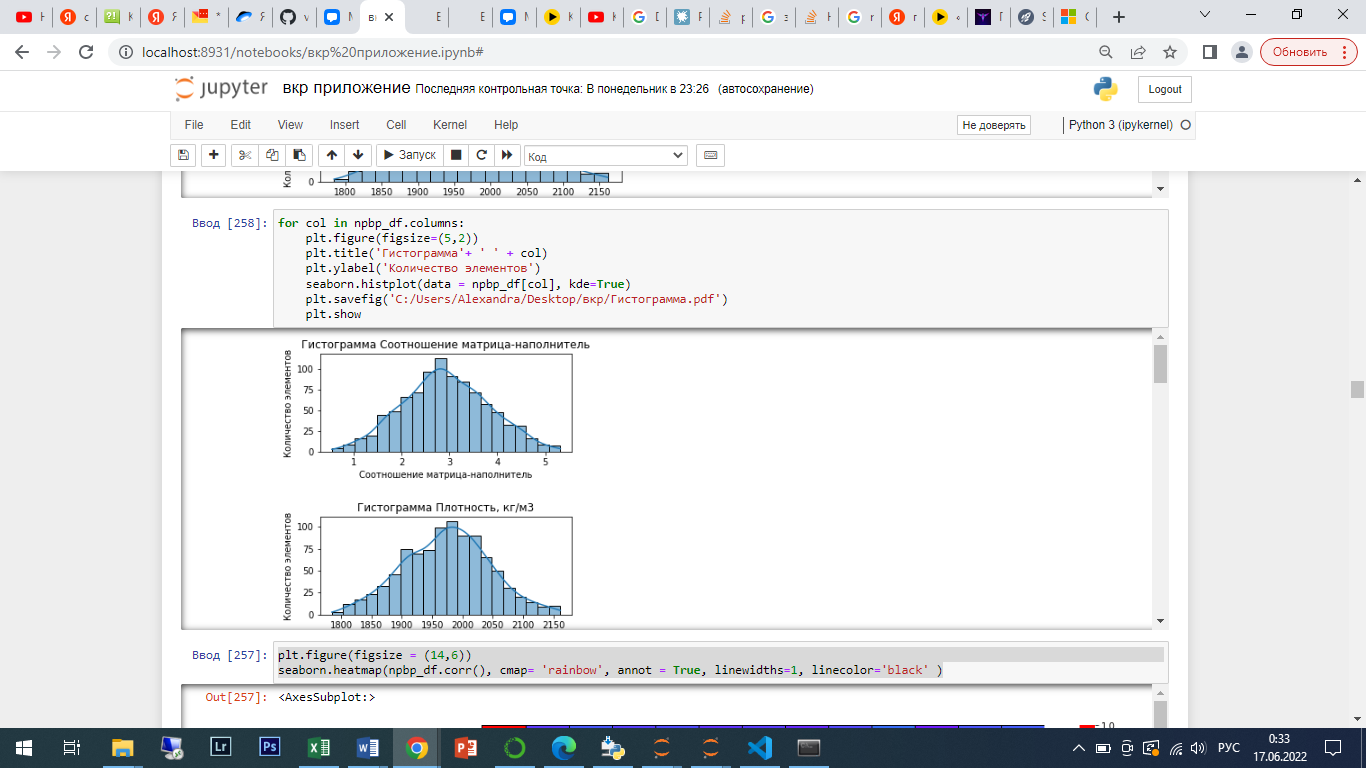
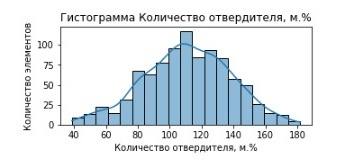
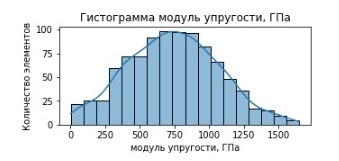


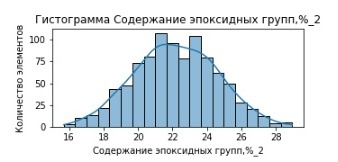
Рисунок 4 – Pairplot данных

1. **Практическая часть**

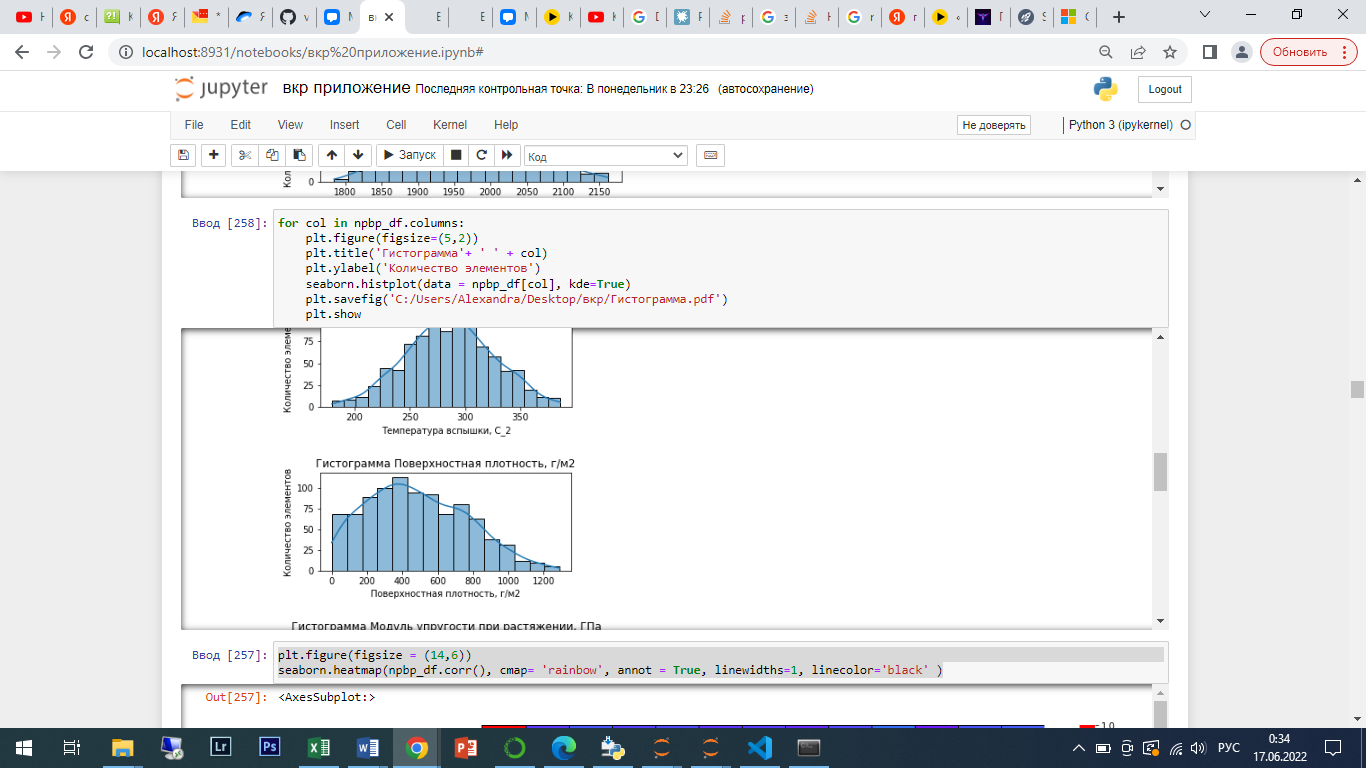
**Предобработка данных**











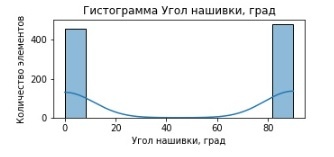
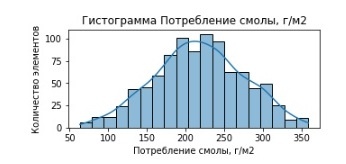
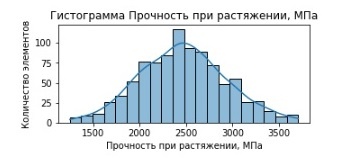
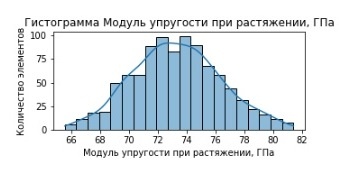




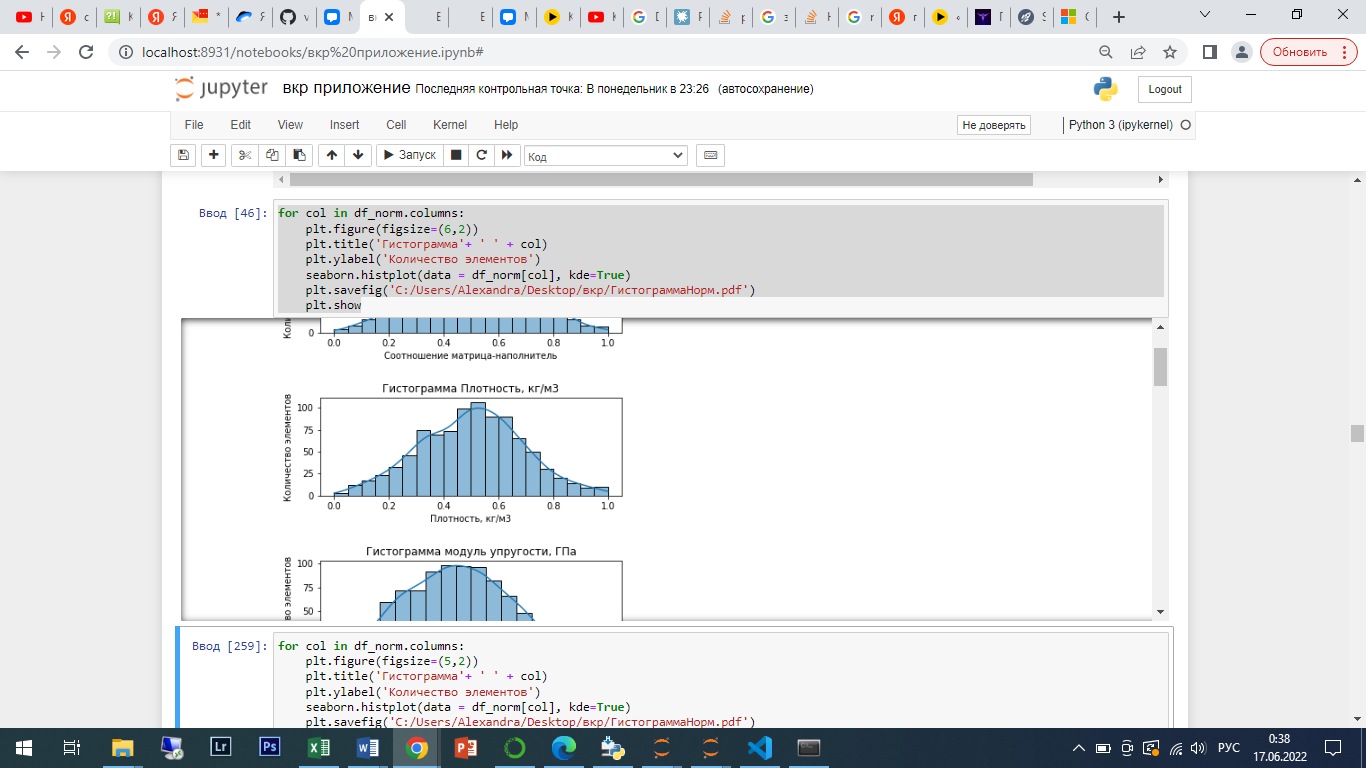


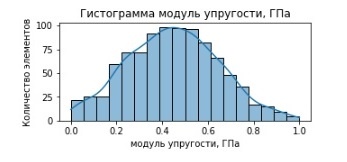
Рисунок 5 – Графики распределения данных без выбросов

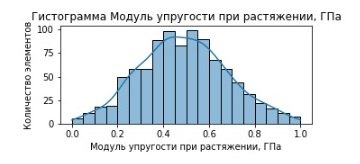
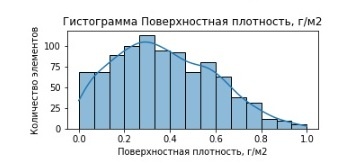
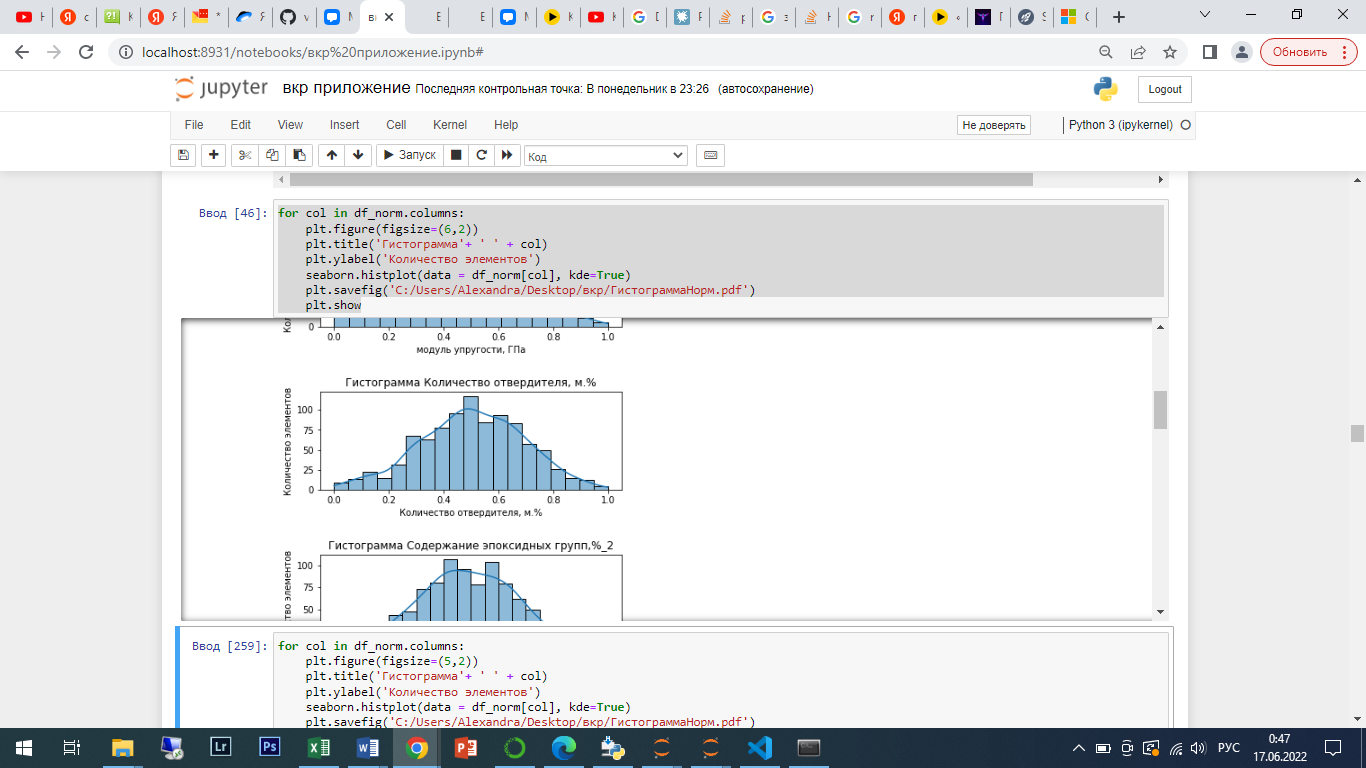


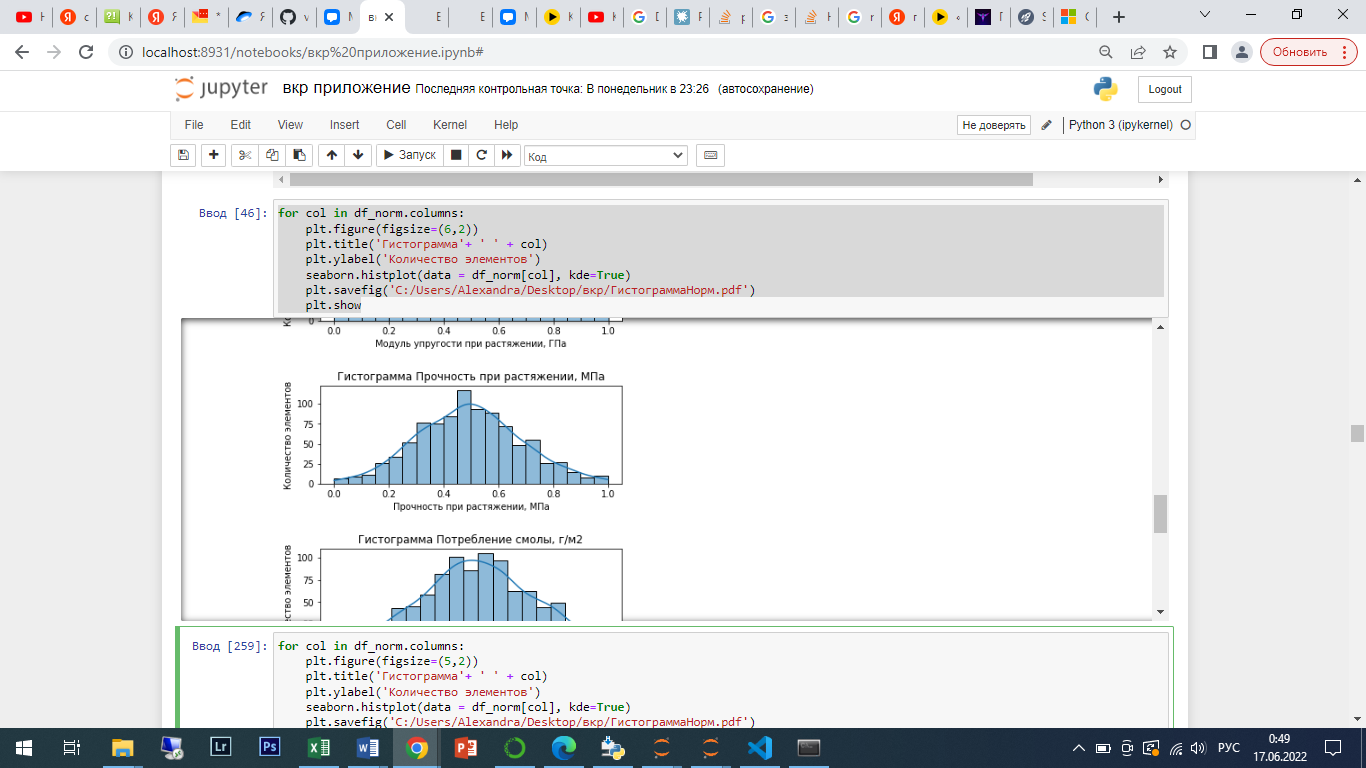
Рисунок 6 – Статистика значений данных без выбросов











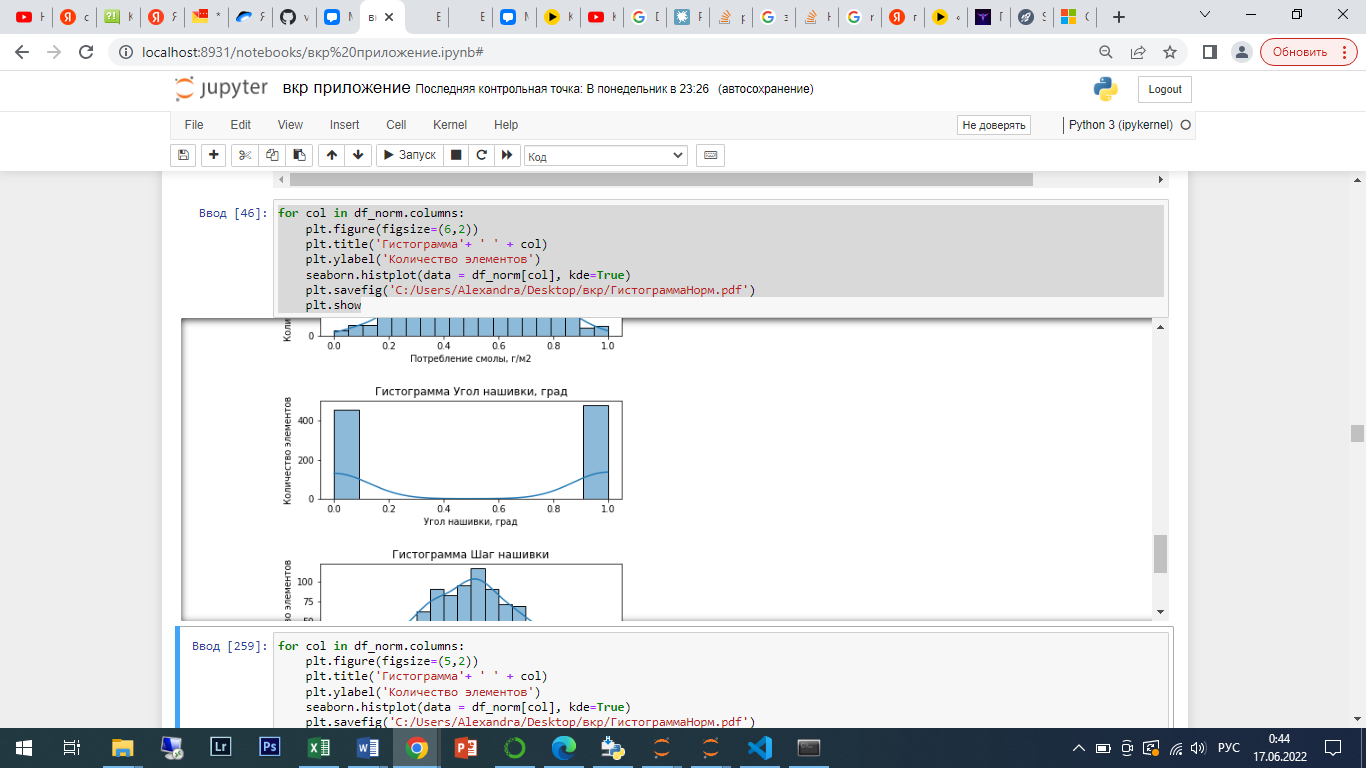




Рисунок 7– Графики распределения нормализированных данных без выбросов

* 1. **Разработка и обучение модели**

При анализе модели был сделан выбор в пользу двух методов:

- Линейная регрессия (LinearRegression)

-Решающие деревья (Decisiontree)

Линейная регрессия - самый простой метод регрессии, который достаточно легок в интерпритации. Регрессия полезна для прогнозированияответа при новых условиях.

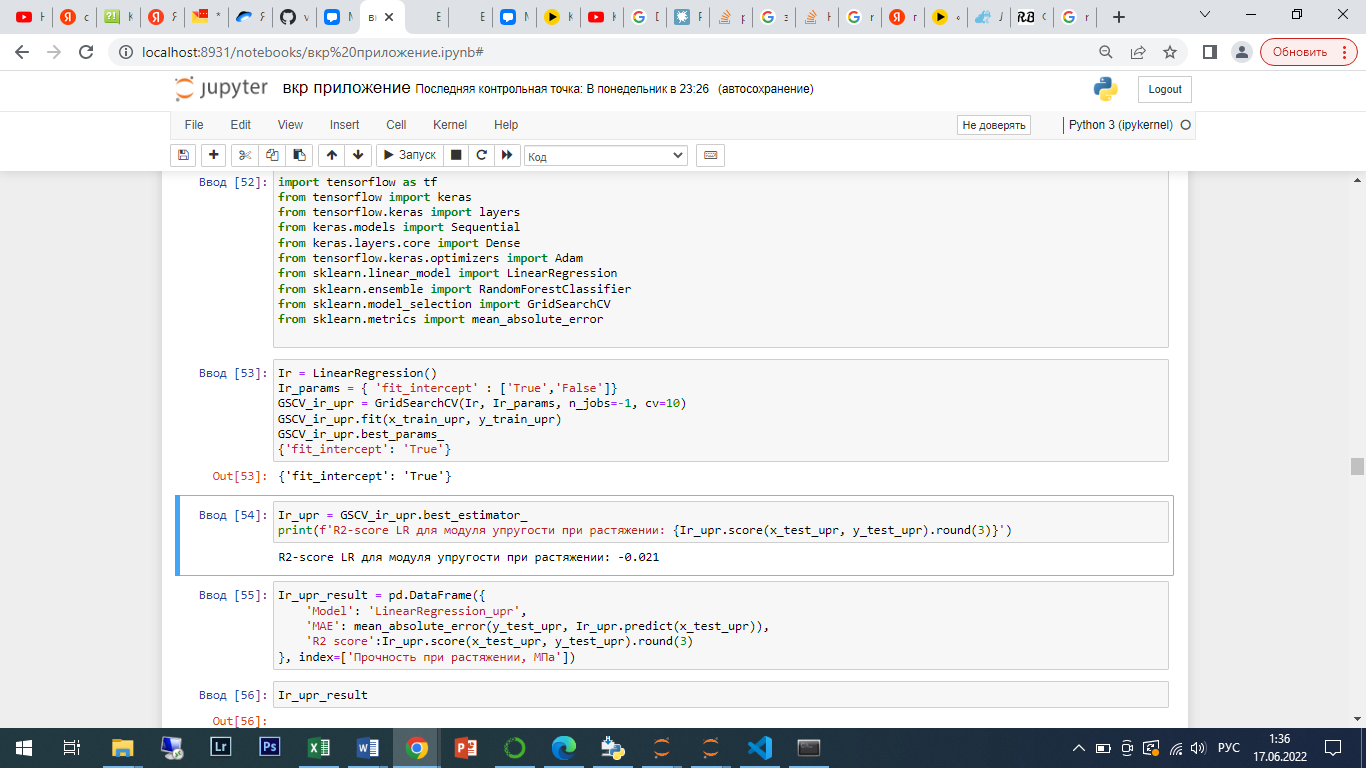


Рисунок 8 – Код обучения модели, основанный на линейной регрессии



Рисунок 9 – Значения модели, основанной на линейной регрессии

При хорошем обучении модели коэффициент детерминации должен стремиться к единице. В данном исследовании значение равно -0.021, что говорит о не самом удачном прогнозе.

Средняя абсолютная ошибка должна стремиться к 0. Данное значение также говорит про отсутствие идеальности в обученной модели.

KNN – достаточно понятный метод, при котором можно получить удовлетворяющий результат, но не увидеть причино-следственную связь. Подходит для количественных нормализованных данных.

Решающие деревья – самый простой метод ансамблей решателей. На данный момент используется не часто, но при проведение исследований было принято решение взять его как эксперементальный вариант и посмотреть результат.

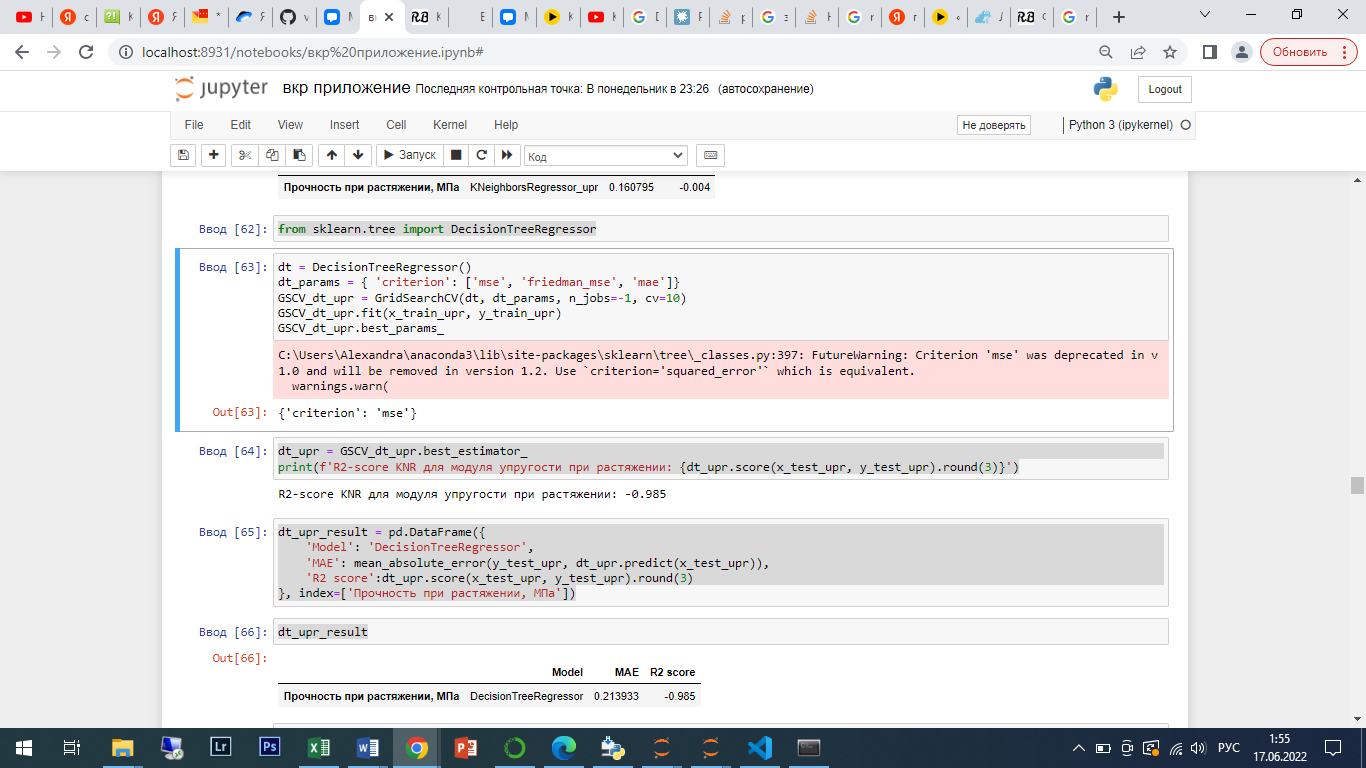


Рисунок 10 – Код обучения модели основанный на классификации по решающим деревьям

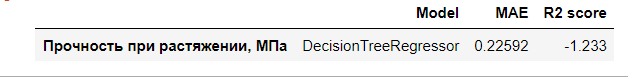


Рисунок 11 – Значения модели, основанной на модели решающих деревьев

Значения не оправдали ожиданий. Из двух созданных моделей эта модель проявила себя наихудшим образом.

* 1. **Тестирование модели**

Для наилучшего восприятия моделей они были визуализированы в графики и показаны в рисунках ниже, где синий цвет является нашей истиной моделью, а зеленый-спрогнозированной:

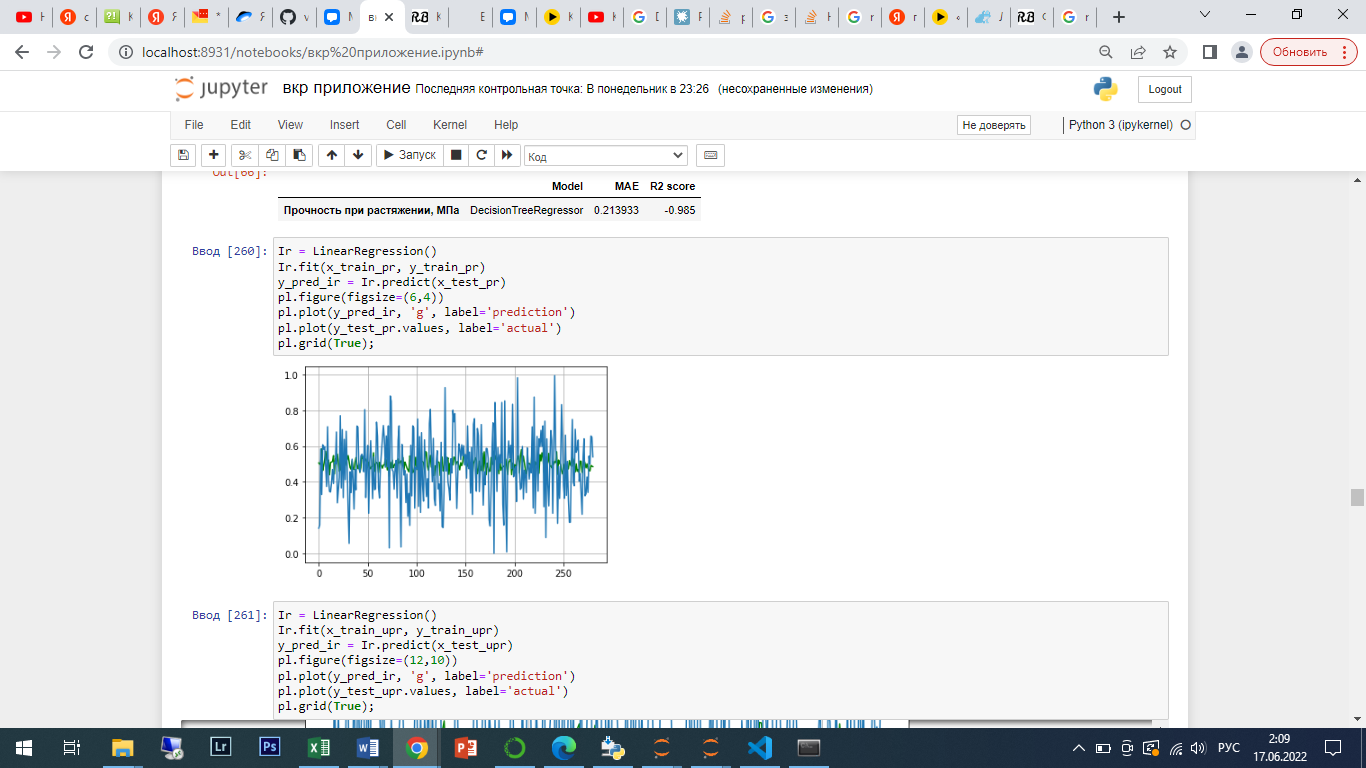


Рисунок 12 – Значения прогноза прочности при растяжении, основанной на модели линейной регрессии

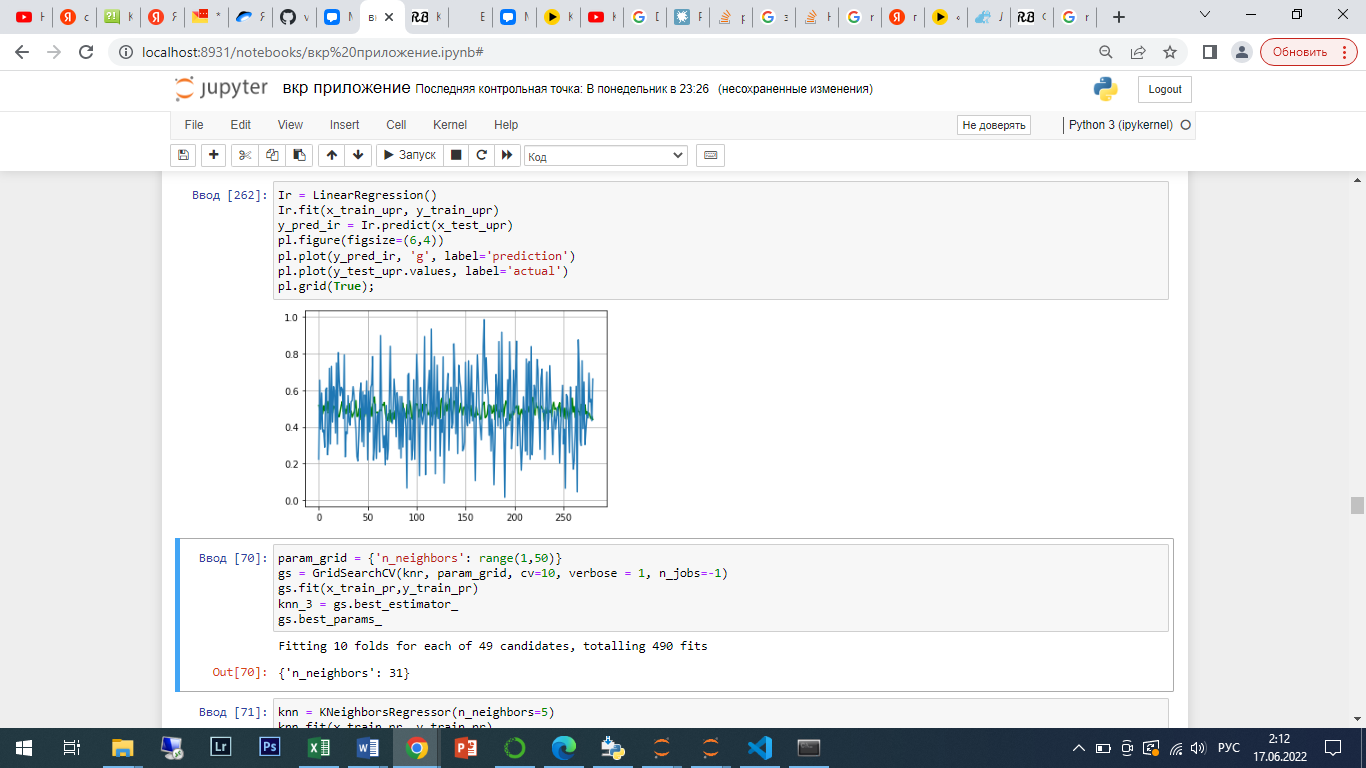


Рисунок 13 – Значения прогноза модуля упругости при растяжении, основанной на модели линейной регрессии

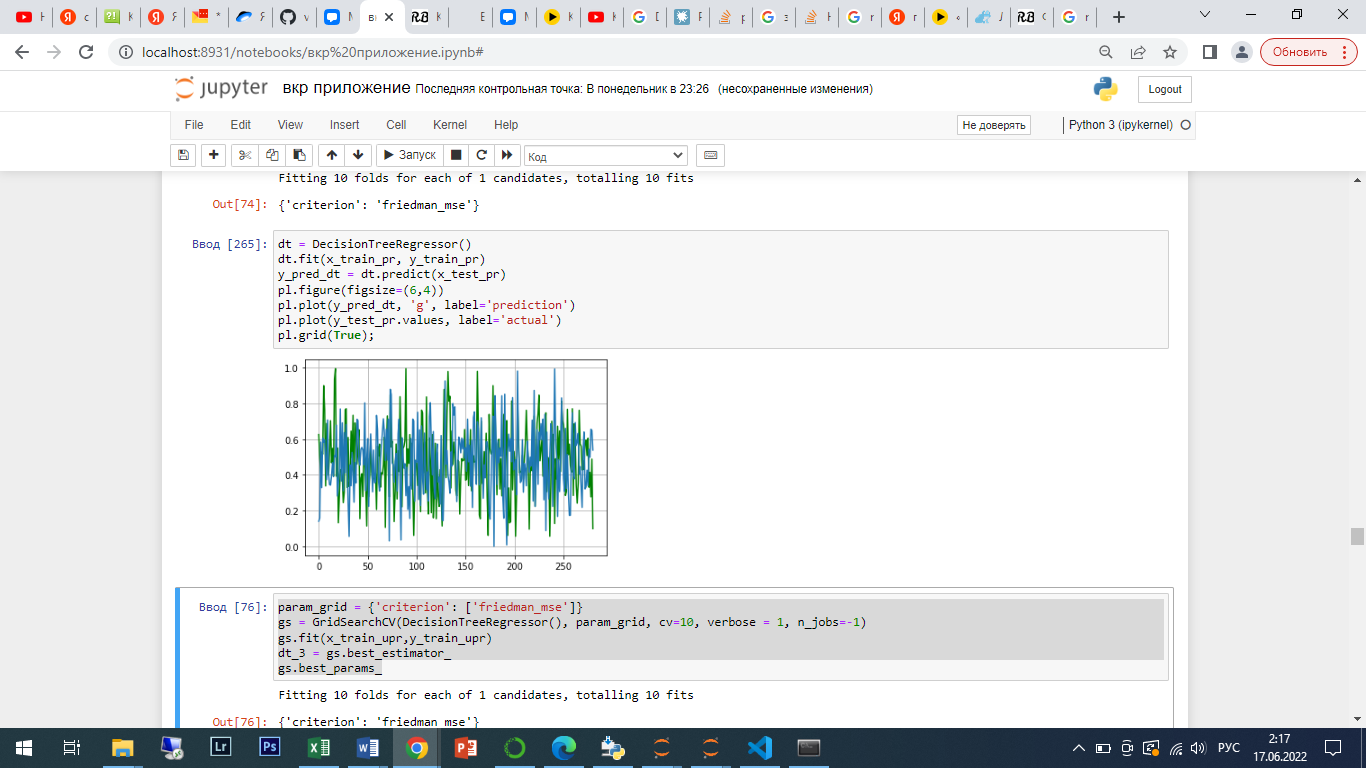


Рисунок 14 – Значения прогноза прочности при растяжении, основанной на модели метода решающих деревьев

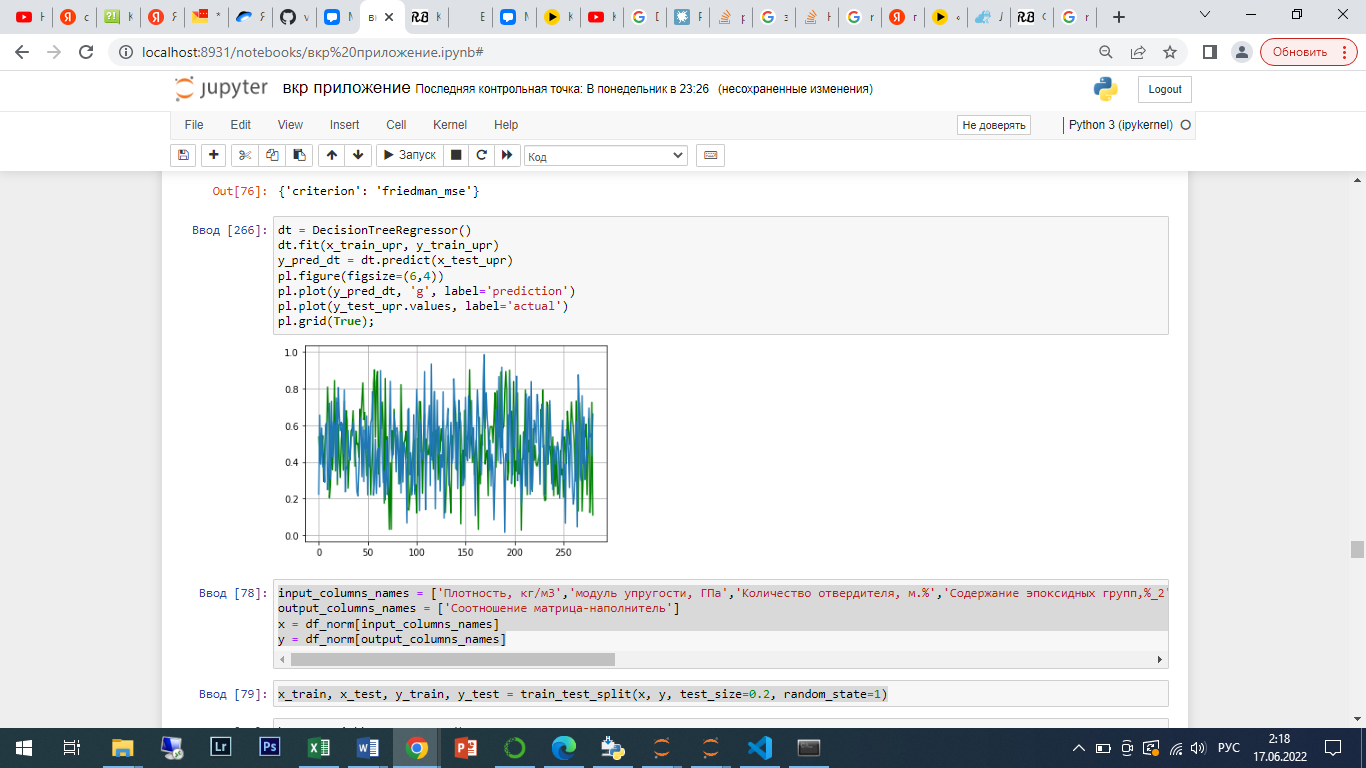


Рисунок 15 – Значения прогноза модуля упругости при растяжении, основанной модели метода решающих деревьев

В результате ни одна из моделей не имеет оценку хорошего прогноза. Данный результат может быть обоснован отсутствием каких-либо зависимостей между значениями. Это могло произойти при огромной погрешности проведенных измерений или при интерполяции большого количества значений.

* 1. **Нейронная сеть**

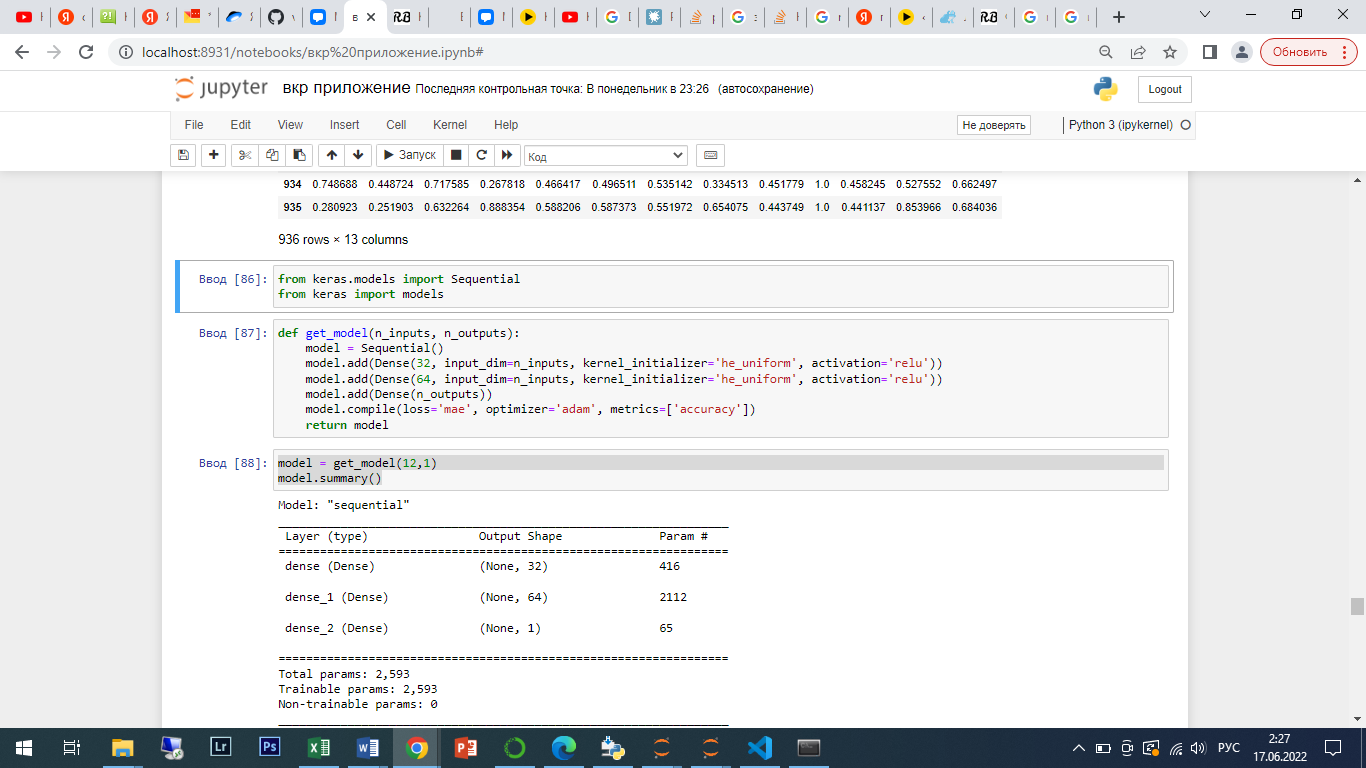


Рисунок 16 – Структура нейронной сети

Модель Sequential– последовательная модель, представляет собой линейный стек слоев. Применяется для моделей, где число скрытых слоев может быть сколько угодно большим.

Имеет 2 полносвязных слоя Denseпо 32 и 64 нейрона, где kernel - это матрица весов, созданная слоем (kernel\_initializer- инициализатор матрицы весов kernel), а activation - установка функции активации для слоя (этом параметре указывается имя встроенной функции или вызываемый объект)

Далее осуществляется выход на 1 нейрон.

Optimizer- определяет процедуру обучения. Передаем в него экземпляры оптимизатора из модуля tf.keras.optimizers, в данном случае им является «Adam»

Происходит компиляция, при параметрах loss- функция, которая минимизируется в процессе обучения (реди распространенных вариантов среднеквадратичная ошибка(mse), categorical\_crossentropy, binary\_crossentropy), metrics- используются для мониторинга обучения (это строковые имена или вызываемые объекты из модуля tf.keras.metrics)

model.summary() - печатает сводное представление об архитектуренейросети.

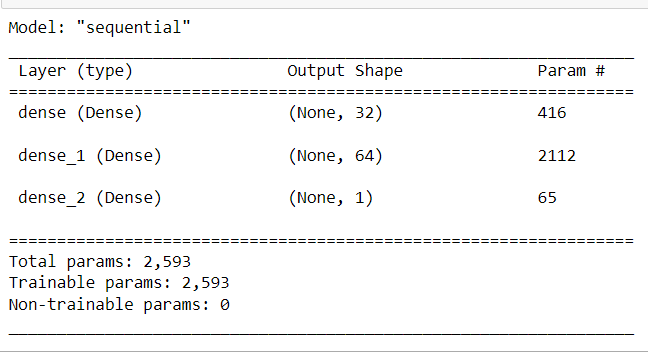


Рисунок 17 – model.summary

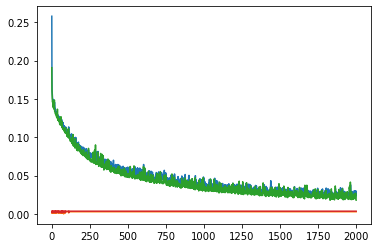


Рисунок 18 – График точности (красный) и ошибки нейросети (зеленый)

Конечное значение средней абсолютной ошибки равно 0.0221, что говорит о хорошем результате, но при этом значение точности перестало расти на отметке 0.0053, что является плохим результатом и говорит о плохом прогнозе значений «Матрица-наполнитель».

* 1. **Создание удаленного репозитория и загрузка результатов работы на него**

Для создания репозитория была выбрана площадка «GitHab».

Имя пользователя – RobinGod80

Название репозитория – VKR

Ссылка - <https://github.com/RobinGod80/VKR>

**Вывод**

В процессе данного задания были получены практические навыки работы с датасетами реальных данных: их преобразование, анализ, предобработка, построение и обучение модели, прогнозирование, а также написание нейронной сети. Были изучены различные обработки и область их применения. Осуществлено знакомство с директорией «GitHab».

Результатом является код, написанный на языке программирования «Python», и само исследование с его конечными значениями.

Однако, практический результат получился хуже ожидаемого. Предположительно из-за интерполяции большого количества значений или из-за огромной погрешности и некачественно снятых измерений.

**Список литературы**

* [Наур, Петер](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%83%D1%80,_%D0%9F%D0%B5%D1%82%D0%B5%D1%80). 1.8. A Basic Principle of Data Science // [Concise Survey of Computer Methods](http://www.naur.com/Conc.Surv.html). — [Lund](https://ru.wikipedia.org/wiki/%D0%9B%D1%83%D0%BD%D0%B4), 1974. — 397 p. — (Studentlitteratur). — [ISBN 91-44-07881-1](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D1%83%D0%B6%D0%B5%D0%B1%D0%BD%D0%B0%D1%8F:%D0%98%D1%81%D1%82%D0%BE%D1%87%D0%BD%D0%B8%D0%BA%D0%B8_%D0%BA%D0%BD%D0%B8%D0%B3/9144078811).
* William S. Cleveland. [Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics](http://cm.bell-labs.com/cm/ms/departments/sia/doc/datascience.pdf) (англ.) // International Statistical Review : Journal. — Willey & Sons, 2001. — Vol. 69, issue 1. — P. 21—26. — [ISSN](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B6%D0%B4%D1%83%D0%BD%D0%B0%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D0%BD%D1%8B%D0%B9_%D1%81%D0%B5%D1%80%D0%B8%D0%B9%D0%BD%D1%8B%D0%B9_%D0%BD%D0%BE%D0%BC%D0%B5%D1%80) [1751-5823](https://www.worldcat.org/search?fq=x0:jrnl&q=n2:1751-5823). — [doi](https://ru.wikipedia.org/wiki/Doi):[10.1111/j.1751-5823.2001.tb00477.x](https://dx.doi.org/10.1111%2Fj.1751-5823.2001.tb00477.x). [Архивировано](https://web.archive.org/web/20120505023607/http:/cm.bell-labs.com/cm/ms/departments/sia/doc/datascience.pdf) 5 мая 2012 года.
* Е. Ю. Журавлёва. [Эпистемический статус цифровых данных в современных научных исследованиях](http://vphil.ru/index2.php?option=com_content&task=view&id=473&pop=1&page=0&Itemid=52). [Вопросы философии](https://ru.wikipedia.org/wiki/%D0%92%D0%BE%D0%BF%D1%80%D0%BE%D1%81%D1%8B_%D1%84%D0%B8%D0%BB%D0%BE%D1%81%D0%BE%D1%84%D0%B8%D0%B8), № 2, 2012, с. 113—123 (1 марта 2012). Дата обращения: 4 декабря 2012.
* Mike Lukides. [What is data science? The future belongs to the companies and people that turn data into products](http://radar.oreilly.com/2010/06/what-is-data-science.html) (англ.). Radar. [O’Reilly Media](https://ru.wikipedia.org/wiki/O%E2%80%99Reilly_Media) (2 июня 2010). Дата обращения: 2 января 2013. [Архивировано](https://www.webcitation.org/6Dy5fWTIa?url=http://radar.oreilly.com/2010/06/what-is-data-science.html) 26 января 2013 года.
* Steve Lohr. [For Today’s Graduate, Just One Word: Statistics](https://www.nytimes.com/2009/08/06/technology/06stats.html?_r=0) (англ.). The New York Times (5 августа 2009). Дата обращения: 2 января 2013. [Архивировано](https://www.webcitation.org/6Dy5gQKLT?url=http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1) 26 января 2013 года.
* F. Jack Smith. [Data Science as An Academic Discipline](http://aurora.gmu.edu/spaceweather/images/DataScienceAsAcademicDiscipline.pdf) (англ.) (недоступная ссылка). Data Science Journal, Volume 5, 19, p. 163—164. CODADA (19 октября 2006). Дата обращения: 2 января 2013. [Архивировано](https://www.webcitation.org/6Dy5h9X14?url=http://aurora.gmu.edu/spaceweather/images/DataScienceAsAcademicDiscipline.pdf) 26 января 2013 года.
* Thomas H. Davenport, D. J. Patil. [Data Scientist: The Sexiest Job of the 21st Century](http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr) (англ.). [Harvard Business Review](https://ru.wikipedia.org/wiki/Harvard_Business_Review). [Harvard University Press](https://ru.wikipedia.org/wiki/%D0%98%D0%B7%D0%B4%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D1%81%D1%82%D0%B2%D0%BE_%D0%93%D0%B0%D1%80%D0%B2%D0%B0%D1%80%D0%B4%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D1%83%D0%BD%D0%B8%D0%B2%D0%B5%D1%80%D1%81%D0%B8%D1%82%D0%B5%D1%82%D0%B0) (1 октября 2012). Дата обращения: 2 января 2013. [Архивировано](https://www.webcitation.org/6Dy5hZZ29?url=http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr) 26 января 2013 года.
* Gil Press. [Data Scientists: The Definition of Sexy](https://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/print/) (англ.). [Forbes](https://ru.wikipedia.org/wiki/Forbes) (27 сентября 2012). Дата обращения: 2 января 2013. [Архивировано](https://www.webcitation.org/6Dy5iDDbj?url=http://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/) 26 января 2013 года.
* Vasant Dhar. Data Science and Prediction (англ.) // [Communications of the ACM](https://ru.wikipedia.org/wiki/Communications_of_the_ACM). — 2013. — Vol. 56, no. 12. — P. 64—73. — [ISSN](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B6%D0%B4%D1%83%D0%BD%D0%B0%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D0%BD%D1%8B%D0%B9_%D1%81%D0%B5%D1%80%D0%B8%D0%B9%D0%BD%D1%8B%D0%B9_%D0%BD%D0%BE%D0%BC%D0%B5%D1%80) [0001-0782](https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0001-0782). — [doi](https://ru.wikipedia.org/wiki/Doi):[10.1145/2500499](https://dx.doi.org/10.1145%2F2500499).
* Леонид Жуков. [Профессия Data scientist](http://www.ospcon.ru/files/media/Zhukov.pdf) (PDF). Материалы конференции «Большие данные в национальной экономике». [Открытые системы](https://ru.wikipedia.org/wiki/%D0%9E%D1%82%D0%BA%D1%80%D1%8B%D1%82%D1%8B%D0%B5_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D1%8B_(%D0%B8%D0%B7%D0%B4%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D1%81%D1%82%D0%B2%D0%BE)) (28 октября 2013). Дата обращения: 31 декабря 2013.