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»

**на тему:**

«Прогнозирование конечных свойств новых материалов

(композиционных материалов)».

Слушатель Старостина Ярослава Константиновна

Москва, 2023

Оглавление

[ВВЕДЕНИЕ 3](#_Toc133364591)

1. Аналитическая часть

[1.1. Постановка задачи 5](#_Toc133364592)

[1.2. Описание используемых методов 9](#_Toc133364593)

[1.3. Разведочный анализ данных 24](#_Toc133364594)

[2. Практическая часть 28](#_Toc133364595)

[2.1. Предобработка данных 28](#_Toc133364596)

[2.2. Разработка и обучение модели 29](#_Toc133364597)

[2.3. Тестирование модели 31](#_Toc133364598)

[2.4. Нейронная сеть, для рекомендации соотношения «матрица – наполнитель». 31](#_Toc133364599)

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

[2.6. Создание удалённого репозитория и загрузка……………………………. 36](#_Toc133364600)

ЗАКЛЮЧЕНИЕ………………………………………………………………………. 37

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ…..……………………………... 38

# **ВВЕДЕНИЕ**

Исследования в области создания конструкционных микро- и нанокомпозитов на полимерной основе относятся к числу критических технологий федерального уровня Российской Федерации. Эти исследования поддерживаются и быстро развиваются в передовых странах мира, а соответствующие разработки находят быстрое и широкое использование в технических приложениях. Композиционные материалы — это материалы, состоящие из двух или более компонентов, нерастворимых друг с другом, с чётко обозначенной границей раздела и сильным взаимодействием по всей зоне контакта. Одним из компонентов композитных материалов является непрерывная фаза, он называется матрица, 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.

Использование композиционных материалов при создании современных аппаратов и устройств потребовало учета их характерных особенностей, таких как: анизотропия жесткости и прочности; вязкоупругие свойства; неоднородность упругих и прочностных параметров; опасность разрушения вдоль поверхностей раздела слоев, определяющих несущую способность конструкции. Решение этой проблемы невозможно без комплексных теоретико-экспериментальных исследований, направленных на выяснение физической картины процессов, протекающих в конструкции и в материале, при предполагаемых эксплуатационных нагрузках. Важным звеном в таких исследованиях является разработка математических моделей рассматриваемых явлений, удовлетворяющих требованиям точности и информативности с одной стороны, экономической и практической применимости к инженерным расчетам — с другой.

**Актуальность:** Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

**Целью** данной работы является прогнозирование конечных свойств новых материалов (композиционных материалов).

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

По заданию на входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

В процессе исследовательской работы были разработаны несколько моделей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также была создана нейронная сеть, которая рекомендует (прогнозирует) соотношение «матрицы - наполнитель». На основе нейронной сети было создано веб - приложение с графическим интерфейсом на фреймворке Flask.

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

Для исследовательской работы были даны 2 файла: X\_bp.xlsx, состоящий из 1023 строки и 11 столбцов, и X\_nup.xlsx, состоящий из 1040 строк и 4 столбцов.

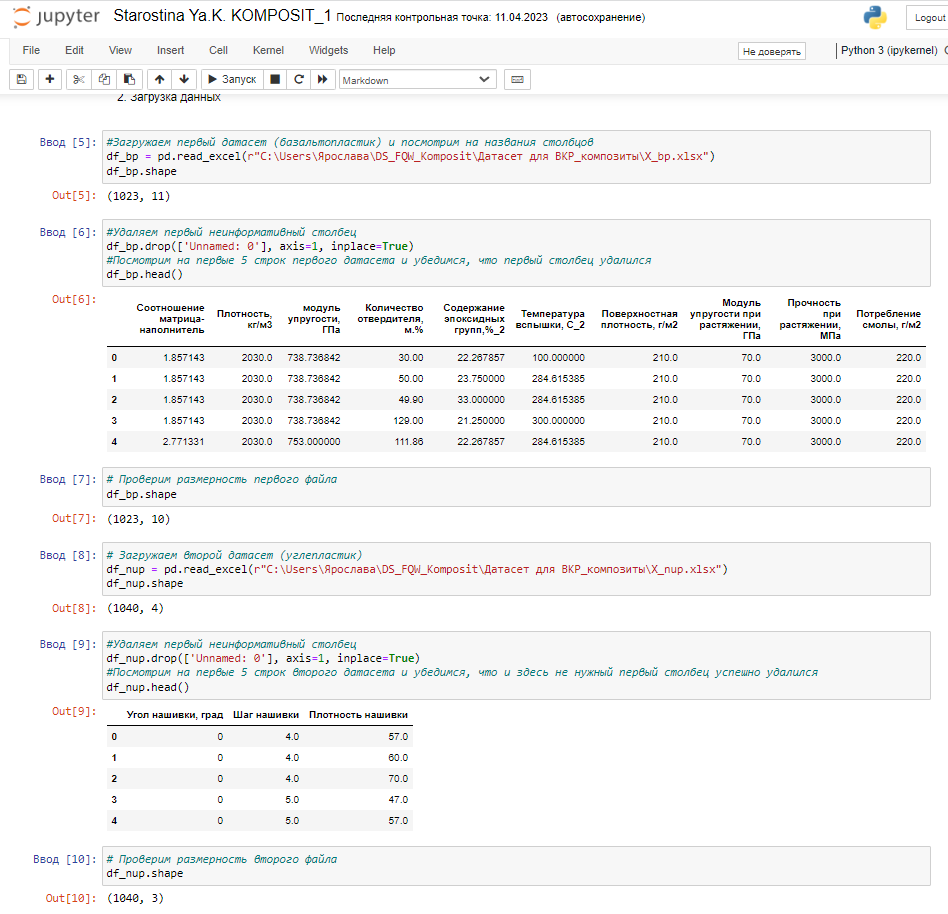


Рисунок 1 – Загрузка файлов

Для решения задачи анализа данных нужно объединить эти два файла. Объединение делать по индексу тип объединения INNER, как указано в задании. Часть информации (17 строк таблицы способов компоновки композитов) не имеют соответствующих строк в таблице соотношений и свойств используемых компонентов композитов, поэтому были удалены.

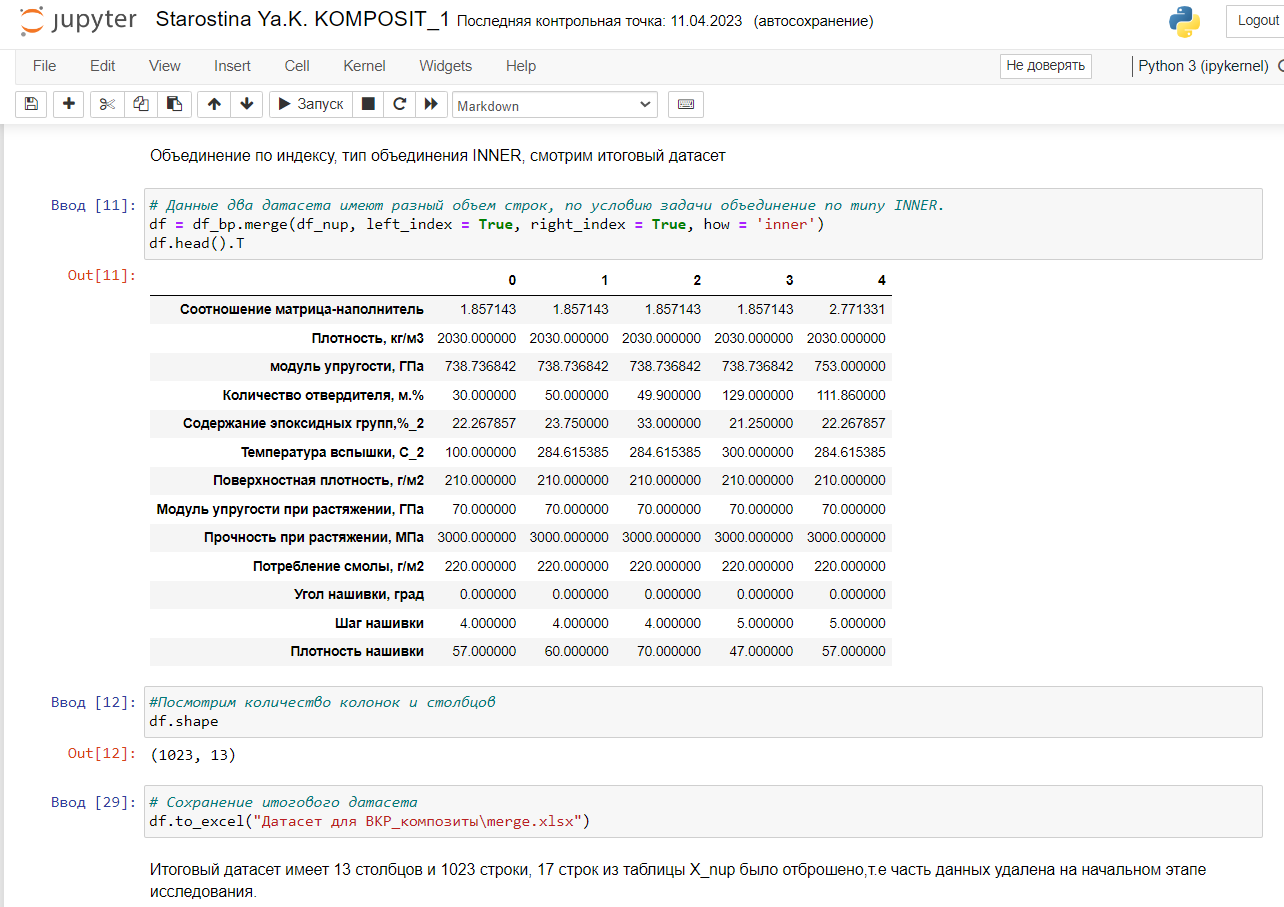


Рисунок 2 – Объединение файлов

Далее проведена оценка данных и характеристика выборки с точки зрения ее особенностей (выбросы, пропуски и т.д.).





Рисунок 3 – Характеристика данных

Все переменные содержат значения float64, качественные характеристики отсутствуют. Пропусков не имеется. Ни одна из записей не является NaN, очистка не требуется. Объединенный файл имеет всего 1023 строки. Принимая во внимание тот факт, что величина "Угол нашивки" является величиной экспериментально исследуемой, поэтому данный столбец не следует подвергать кодировке, так как "в продакшене" можно будет столкнуться и с другими значениями градусов.

Пропуски в данных отсутствуют. Значит, сразу приступаем к работе с выбросами. Выбросы являются точками данных, которые лежат вдали от обычного распределения данных и приводит к тому, что ниже воздействие на общее распределение данных:

* Влияет на общую стандартную вариацию данных.
* Манипулирует общее среднее значение данных.
* Преобразует данные в перекошенную форму.
* Это вызывает смещение в оценке точности модели обучения машины.
* Влияет на распределение и статистику набора данных.

Из-за вышеуказанных причин необходимо обнаружить и избавиться от выбросов до моделирования набора данных. Для удаления выбросов существует 2 основных метода: метод 3-х сигм и метод межквартильных расстояний.

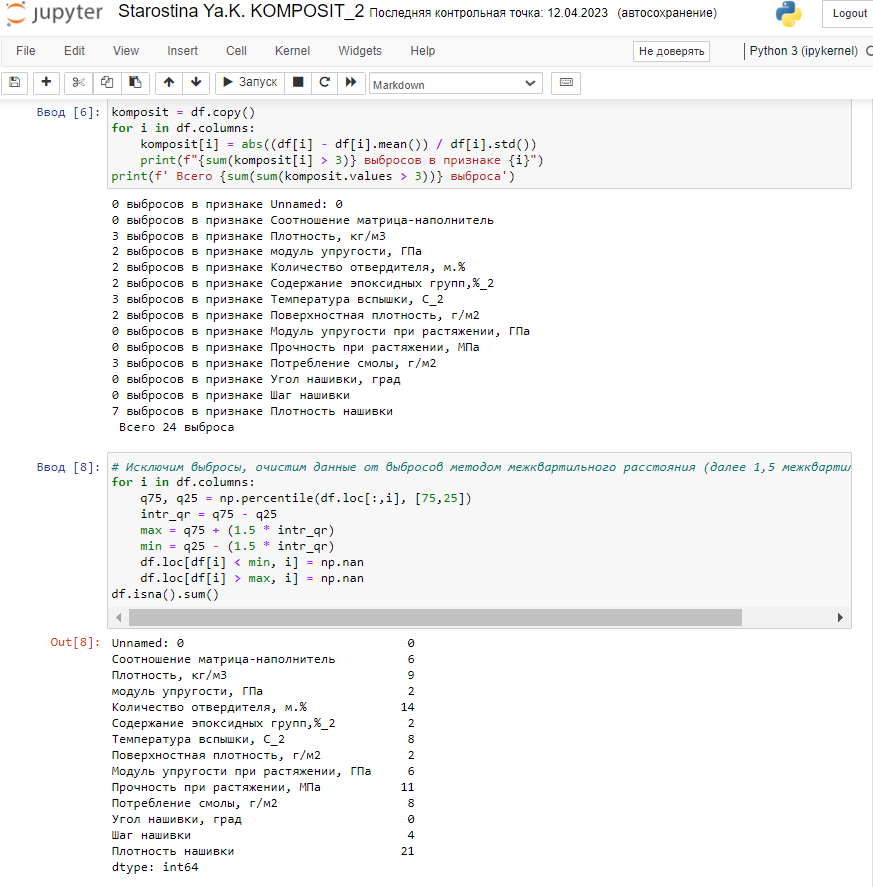


Рисунок 4 – Работа от выбросов в данных

Далее была проведена замены пустых значений на медиану.

* 1. **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:

* Метод линейной регрессии - Linear Regression;
* Метод опорных векторов - Support Vector Regression;
* Метод К ближайших соседей - K Neighbors Regressor;
* Дерево решений - Decision Tree Regressor;
* Метод случайного леса - Random Forest Regressor;
* Модель градиентного спуска - Gradient Boosting Regressor;
* Стохастический градиентный спуск - Stochastic Gradient Descent Regressor;
* Модель Лассо (регрессия с наименьшим углом) - LassoLars;
* Модель регрессия байесовского гребня - Bayesian Ridge;
* Линейная модель с распределением Пуассона - Poisson Regressor;
* Многослойный перцептрон - Multi-layer Perceptron regressor

Приведем краткое описание методов, которые использовались для решения поставленной задачи. Обратим внимание на достоинства, недостатки и области применения каждого из методов, также приведём итоговую сравнительную таблицу с оценками работоспособности каждого метода.

* + 1. **Метод линейной регрессии - Linear Regression**

Линейная регрессия (Linear regression) — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Результаты применения метода представлены на рисунке 5. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R2 , или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R2 к единице, тем лучше.

Достоинства метода: быстр и прост в реализации; легко интерпретируем; имеет меньшую сложность по сравнению с другими алгоритмами;

Недостатки метода: моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

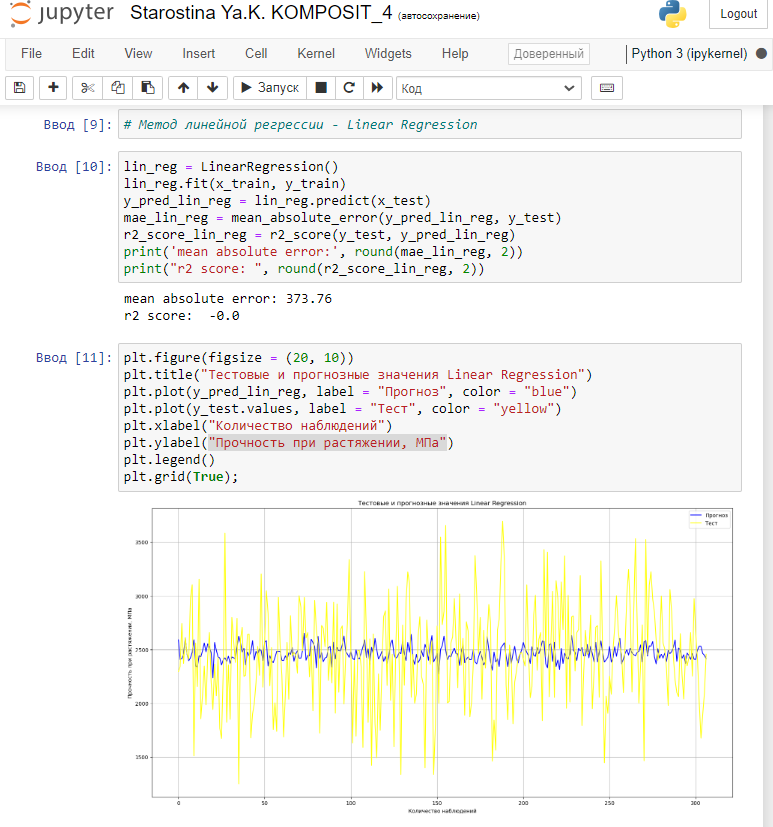
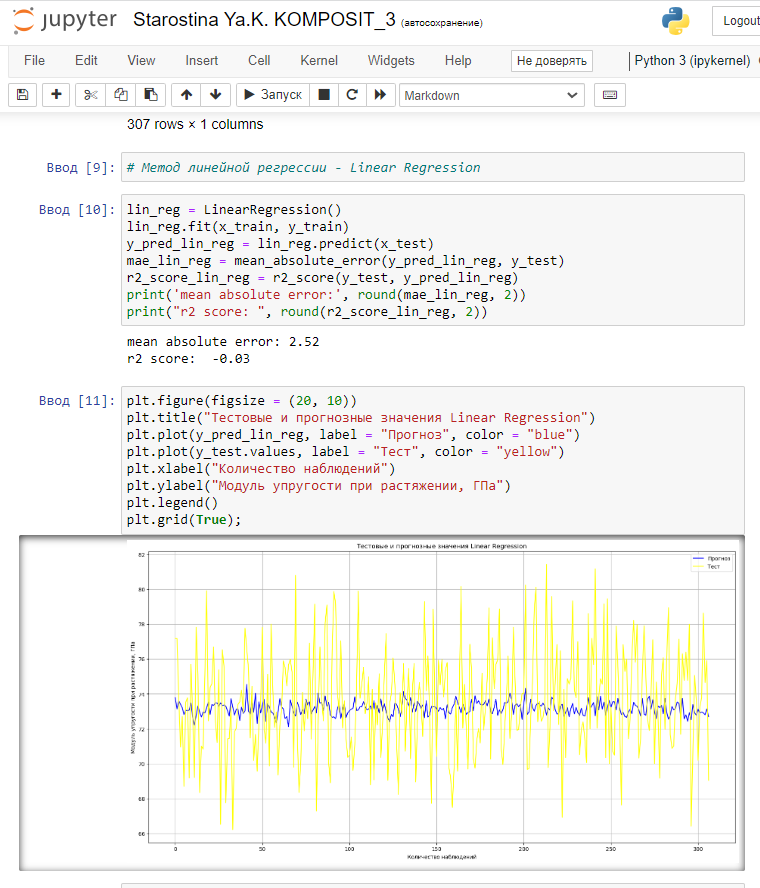


Рисунок 5 – Результаты применения метода линейной регрессии для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Метод опорных векторов - Support Vector Regression**

Метод опорных векторов (Support Vector Regression) – этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших наборах данных. Результаты применения метода представлены на рисунке 6. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использование схожих алгоритмов для анализа данных и распознавания шаблонов. Учитывая обучающую выборку, где алгоритм помечает каждый объект, как принадлежащий к одной из двух категорий, строит модель, которая определяет новые наблюдения в одну из категорий. Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв, и он максимален. Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию.  Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

Недостатки метода: неустойчивость к шуму, поэтому в работе была проведена тщательнейшая работа с выбросами, иначе в обучающих данных шумы становятся опорными объектами-нарушителями и напрямую влияют на построение разделяющей гиперплоскости; для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать, поэтому были рассмотрены и другие методы.

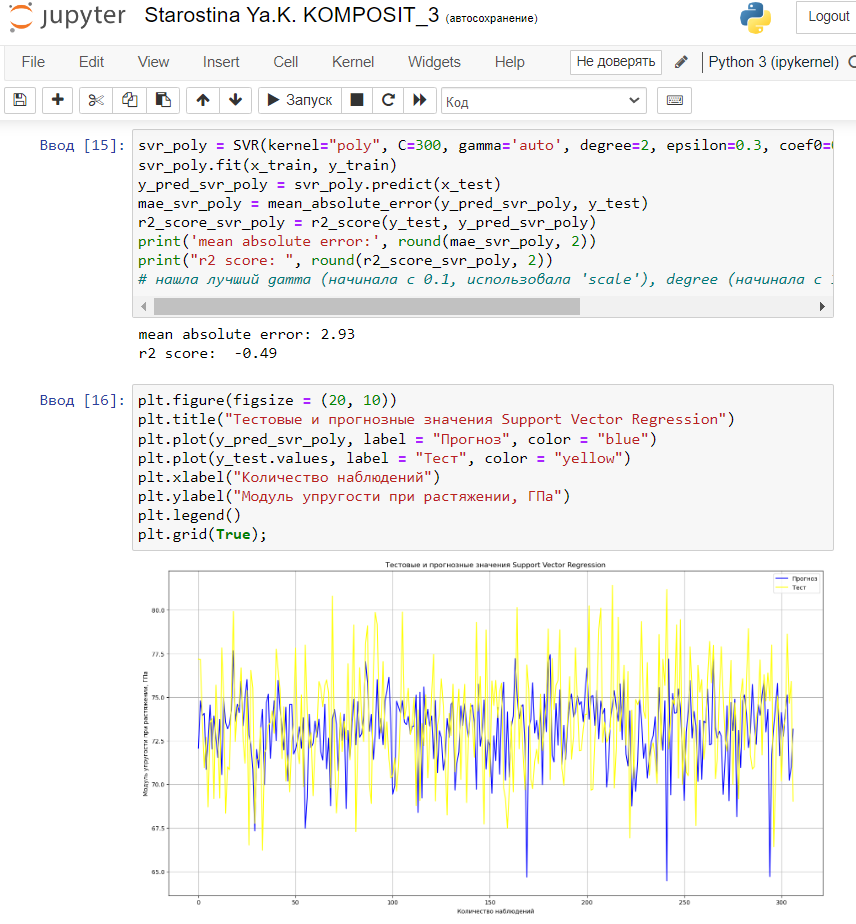
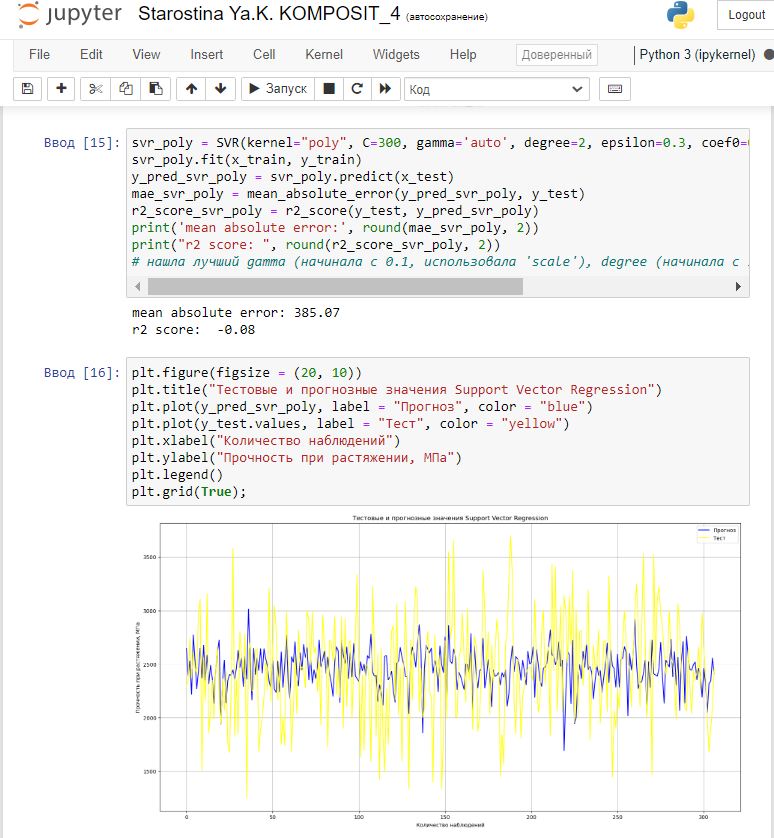
 

Рисунок 6 – Результаты применения метода опорных векторов для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Метод К ближайших соседей - K Neighbors Regressor**

Метод 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. Результаты применения метода представлены на рисунке 7. 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.

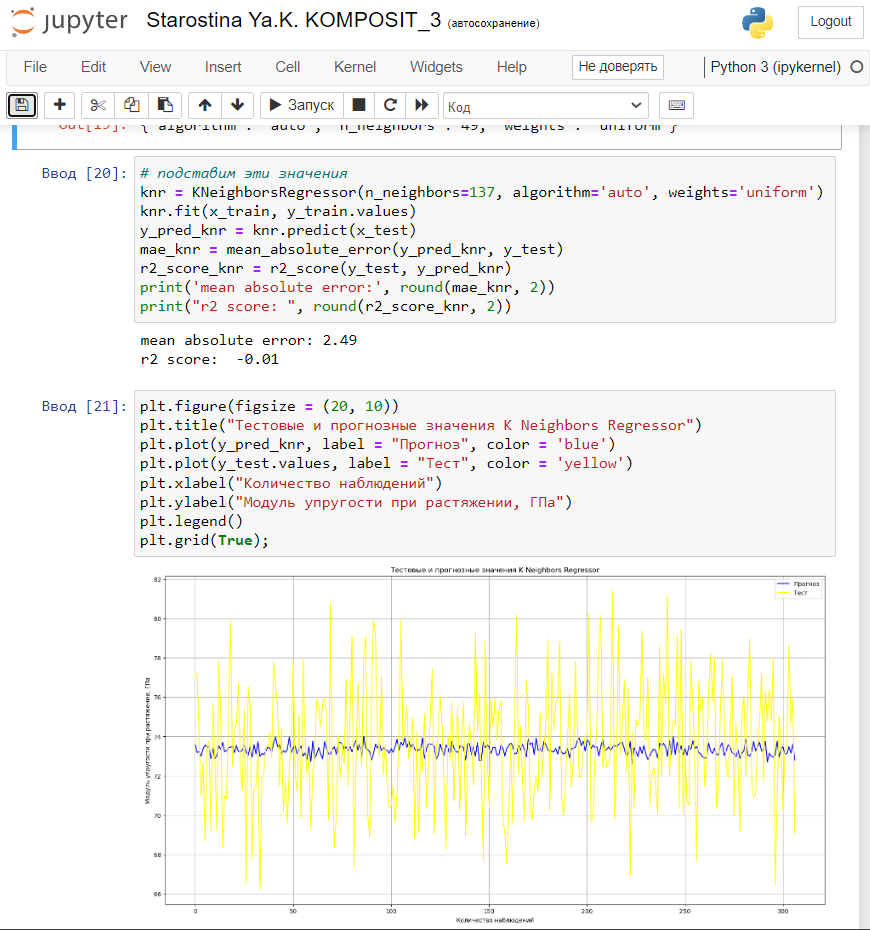
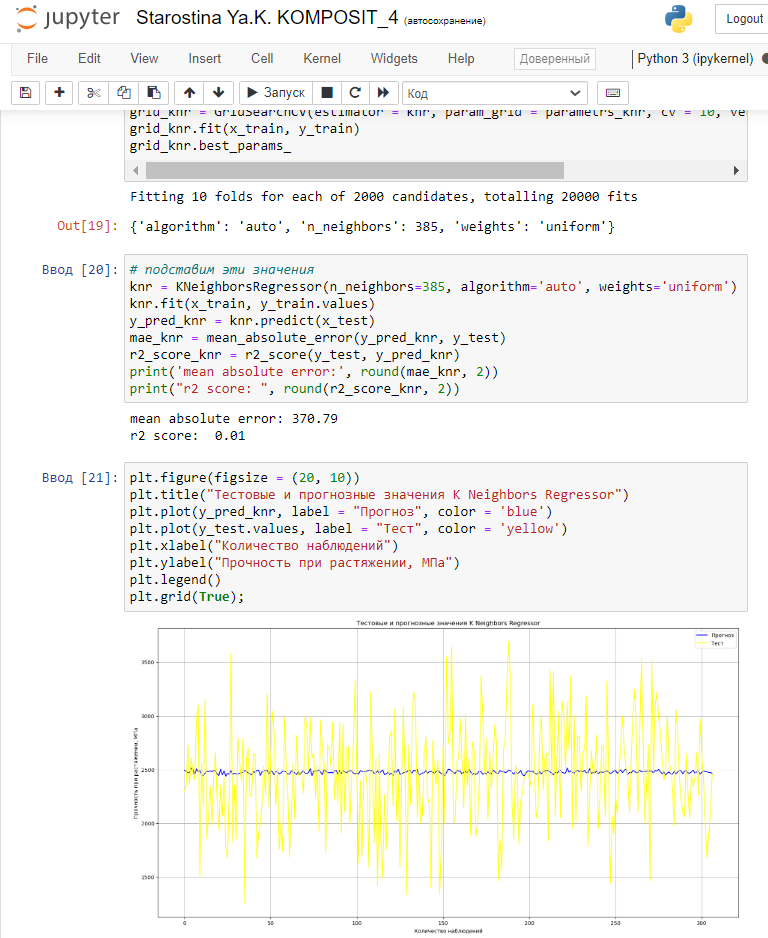
 

Рисунок 7 – Результаты применения метода К ближайших соседей для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Дерево решений - Decision Tree Regressor**

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. Результаты применения метода представлены на рисунке 8. 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.  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.

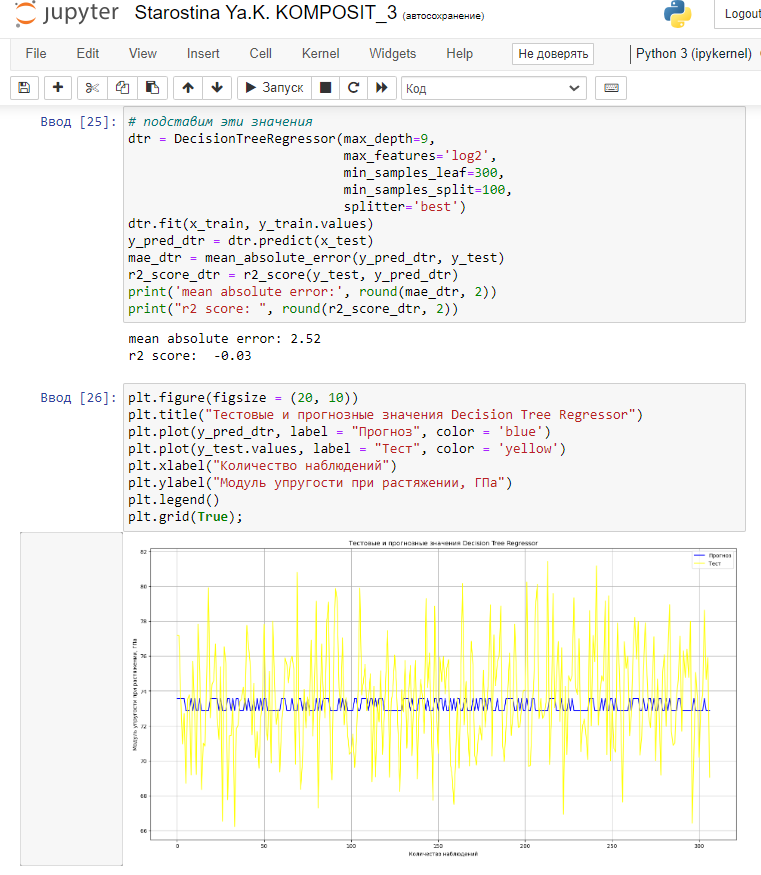
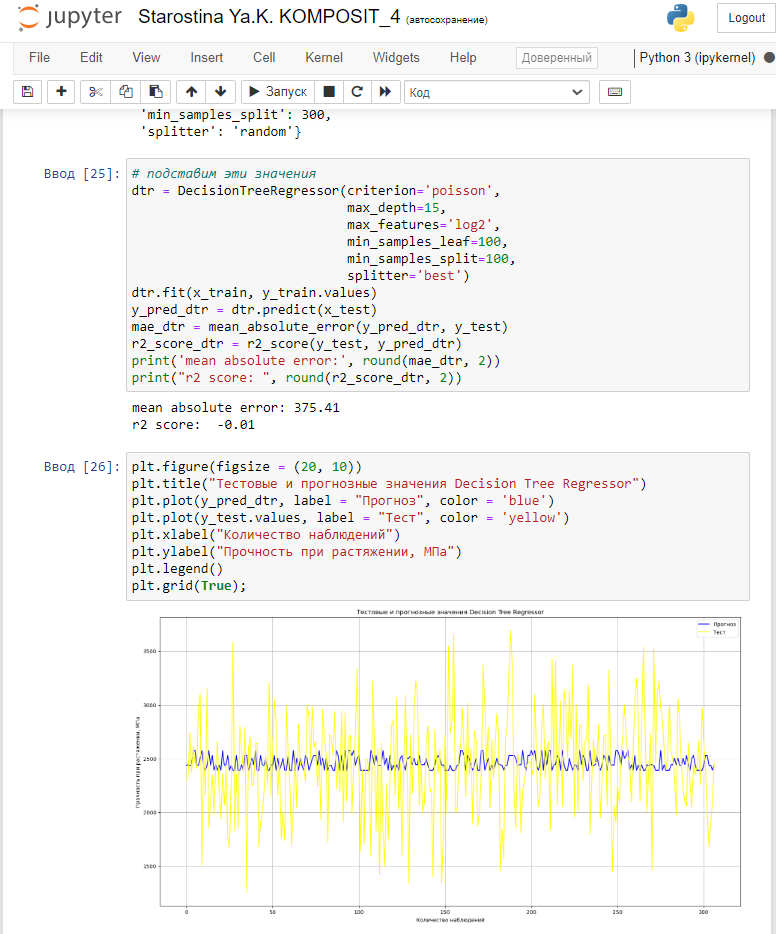
 

Рисунок 8 – Результаты применения метода Decision Tree Regressor для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Метод случайного леса - Random Forest Regressor**

Метод случайного дерева (Random Forest) – это универсальный алгоритм машинного обучения с учителем. Его можно использовать во множестве задач, но в основном он применяется в проблемах классификации и регрессии. Результаты применения метода представлены на рисунке 9. Алгоритм случайного леса (Random Forest) — универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается. Также это один из немногих алгоритмов, который можно использовать в абсолютном большинстве задач.

Достоинства: имеет высокую точность предсказания, которая сравнима с результатами градиентного бустинга; не требует тщательной настройки параметров, хорошо работает из коробки; практически не чувствителен к выбросам в данных из-за случайного семплирования (random sample); не чувствителен к масштабированию и к другим монотонным преобразованиям значений признаков; редко переобучается; в случае наличия проблемы переобучения, она преодолевается путем усреднения или объединения результатов различных деревьев решений; способен эффективно обрабатывать данные с большим числом признаков и классов; хорошо работает с пропущенными данными – сохраняет хорошую точность даже при их наличии; одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки; высокая параллелизуемость и масштабируемость.

Недостатки: для реализации алгоритма случайного дерева требуется значительный объем вычислительных ресурсов; большой размер моделей; построение случайного леса отнимает больше времени, чем деревья решений или линейные алгоритмы; алгоритм склонен к переобучению на зашумленных данных; нет формальных выводов, таких как p-values, которые используются для оценки важности переменных; в отличие от более простых алгоритмов, результаты случайного леса сложнее интерпретировать; когда в выборке очень много разреженных признаков, таких как тексты или наборы слов (bag of words), алгоритм работает хуже чем линейные методы; в отличие от линейной регрессии, Random Forest не обладает возможностью [экстраполяции](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D1%81%D1%82%D1%80%D0%B0%D0%BF%D0%BE%D0%BB%D1%8F%D1%86%D0%B8%D1%8F). Это можно считать и плюсом, так как в случае выбросов не будет экстремальных значений; если данные содержат группы признаков с корреляцией, которые имеют схожую значимость для меток, то предпочтение отдается небольшим группам перед большими, что ведет к недообучению; процесс прогнозирования с использованием случайных лесов очень трудоемкий по сравнению с другими алгоритмами.

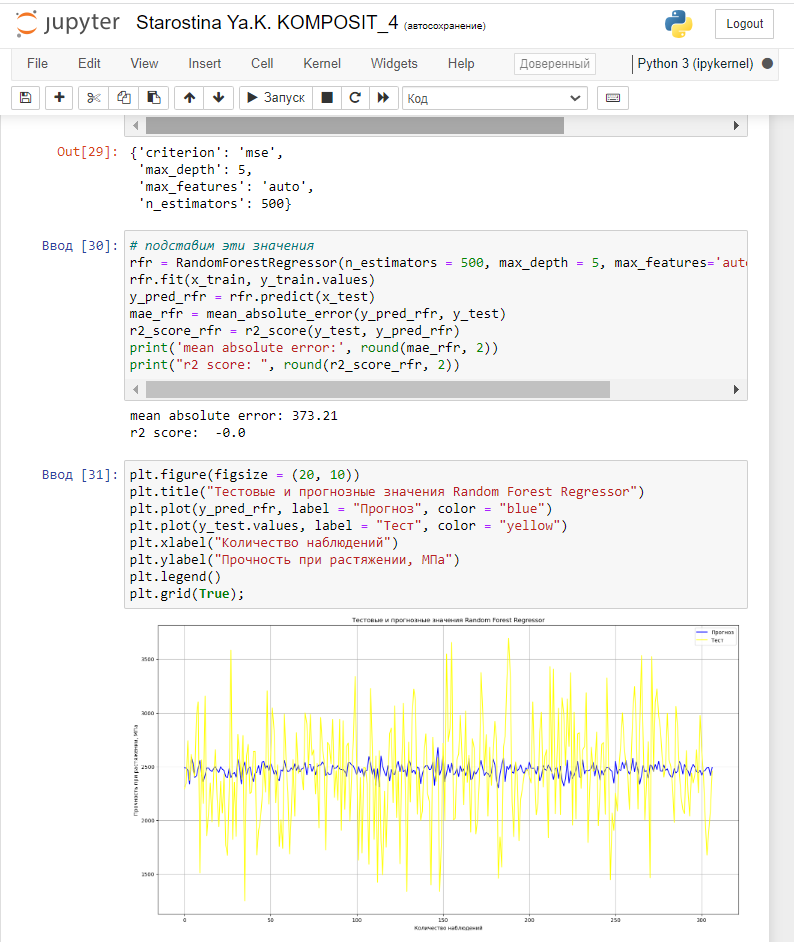
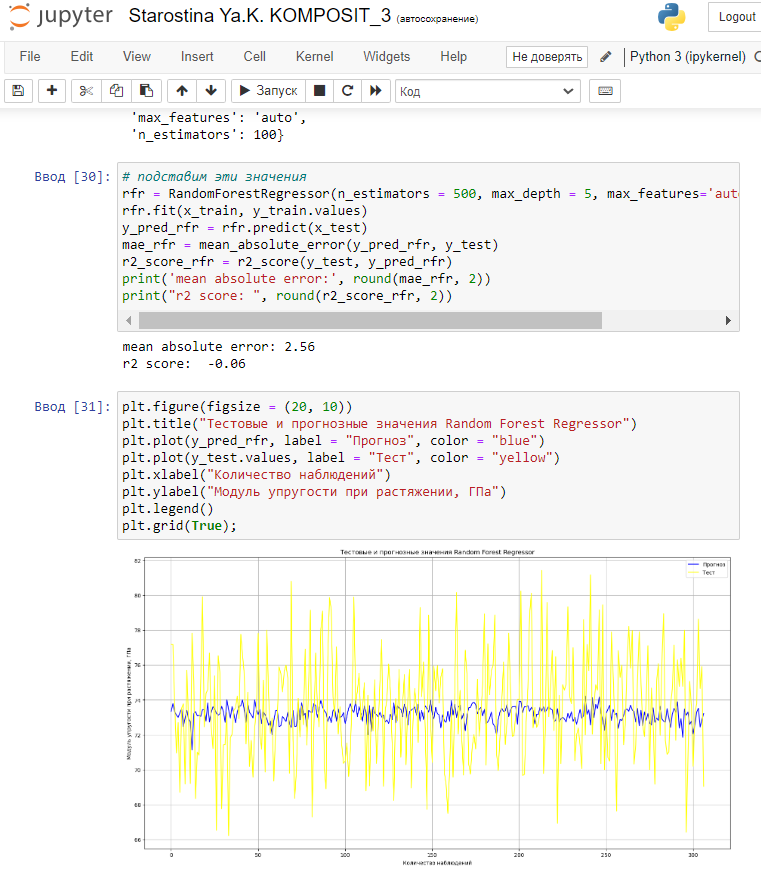


Рисунок 9 – Результаты применения модели случайного леса для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Модель градиентного бустинга - Gradient Boosting Regressor**

Модель градиентного бустинга (Gradient Boosting Regressor) — это ансамбль деревьев решений, обученный с использованием градиентного бустинга. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Результаты применения метода представлены на рисунке 10. Основная идея градиентного бустинга: строятся последовательно несколько базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.

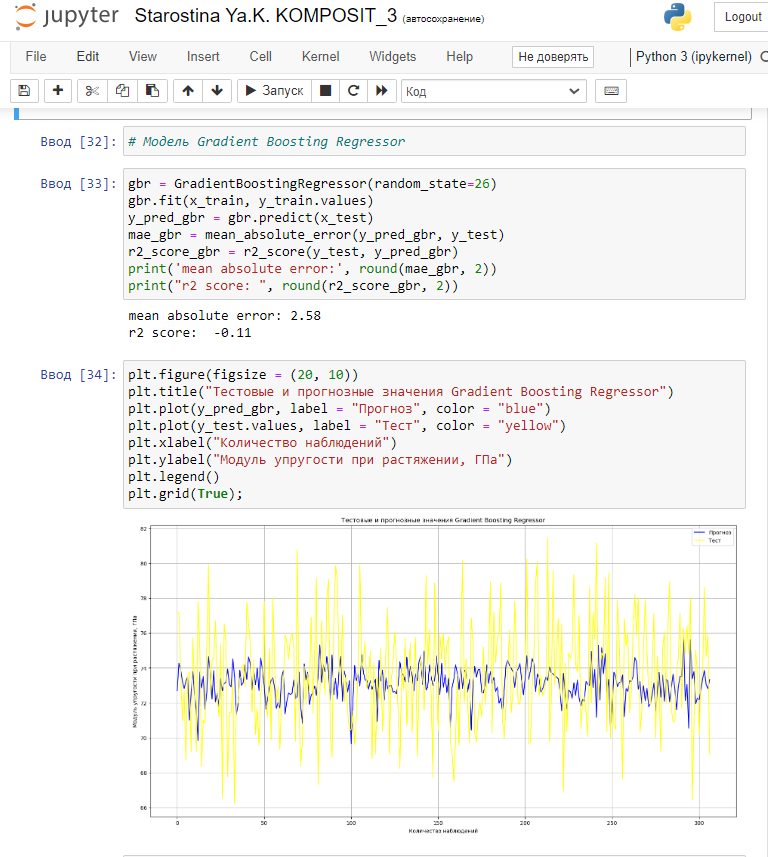
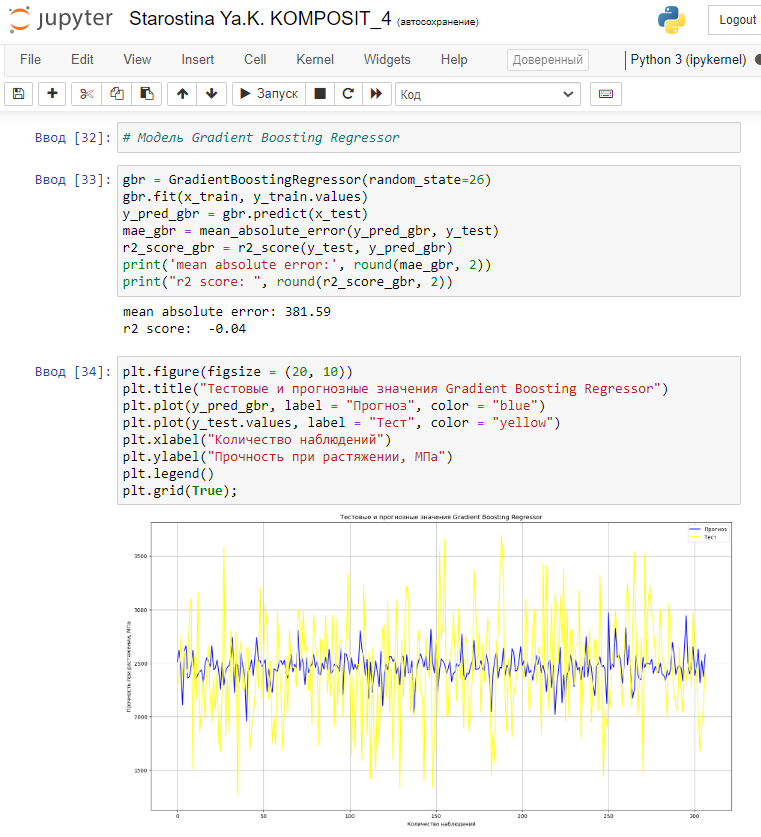
 

Рисунок 10 – Результаты применения модели градиентного бустинга для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Стохастический градиентный спуск - Stochastic Gradient Descent Regressor**

Стохастический градиентный спуск (SGDRegressor) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь. Этот подход подразумевает корректировку весов нейронной сети, используя аппроксимацию градиента функционала, вычисленную только на одном случайном обучающем примере из выборки. Результаты применения метода представлены на рисунке 11.

Достоинства метода: эффективен; прост в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках.

Недостатки метода: требует ряд гиперпараметров; чувствителен к масштабированию функций; может не сходиться или сходиться слишком медленно; функционал многоэкстремален; процесс может "застрять" в одном из локальных минимумов; возможно переобучение.

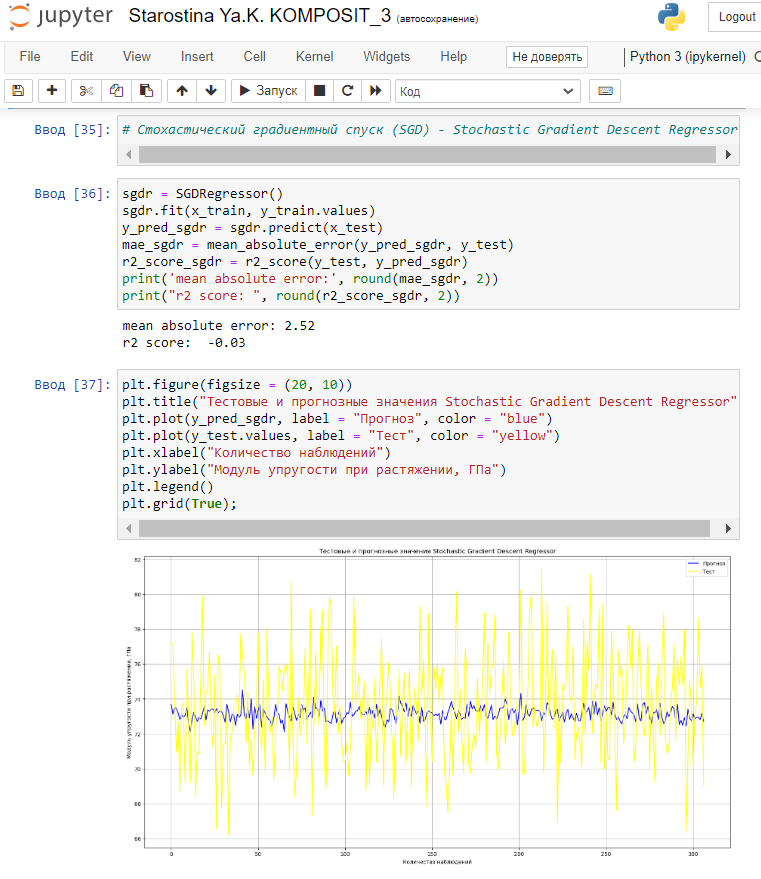
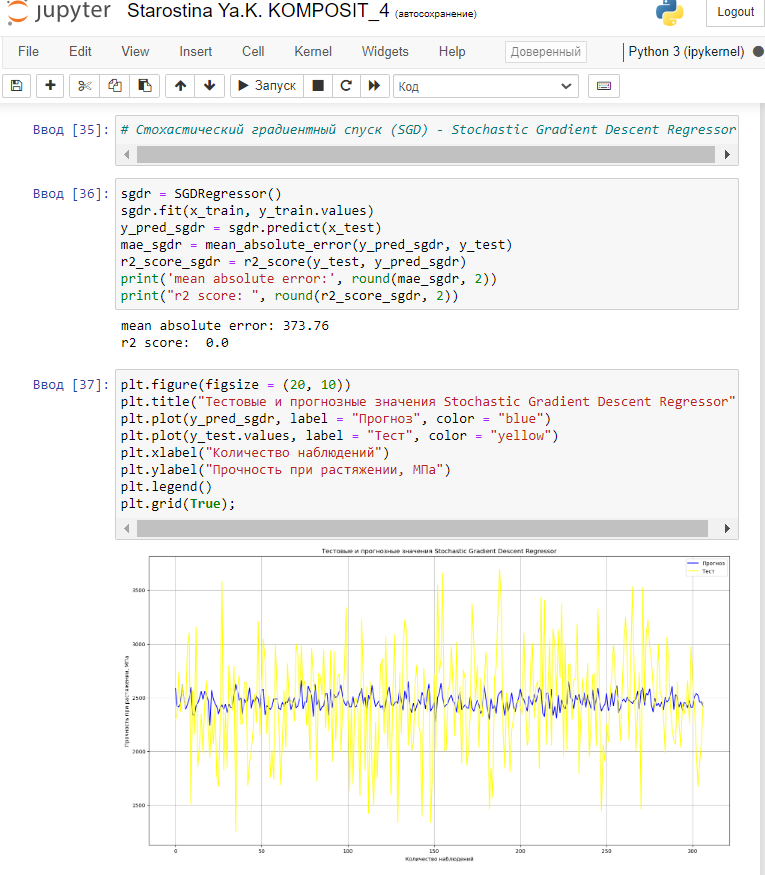
 

Рисунок 11 – Результаты применения модели стохастический градиентного спуска для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Модель Лассо (регрессия с наименьшим углом) – LassoLars**

Лассо регрессия (Lasso) — это линейная модель, которая оценивает разреженные коэффициенты.  Это простой метод, позволяющий уменьшить сложность модели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Результаты применения метода представлены на рисунке 12. Это даёт более устойчивое решение. В регрессии лассо добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квадратичного смещения, используется смещение абсолютного значения. Лассо регрессия хорошо прогнозирует модели временных рядов на основе регрессии, таким как авторегрессии.

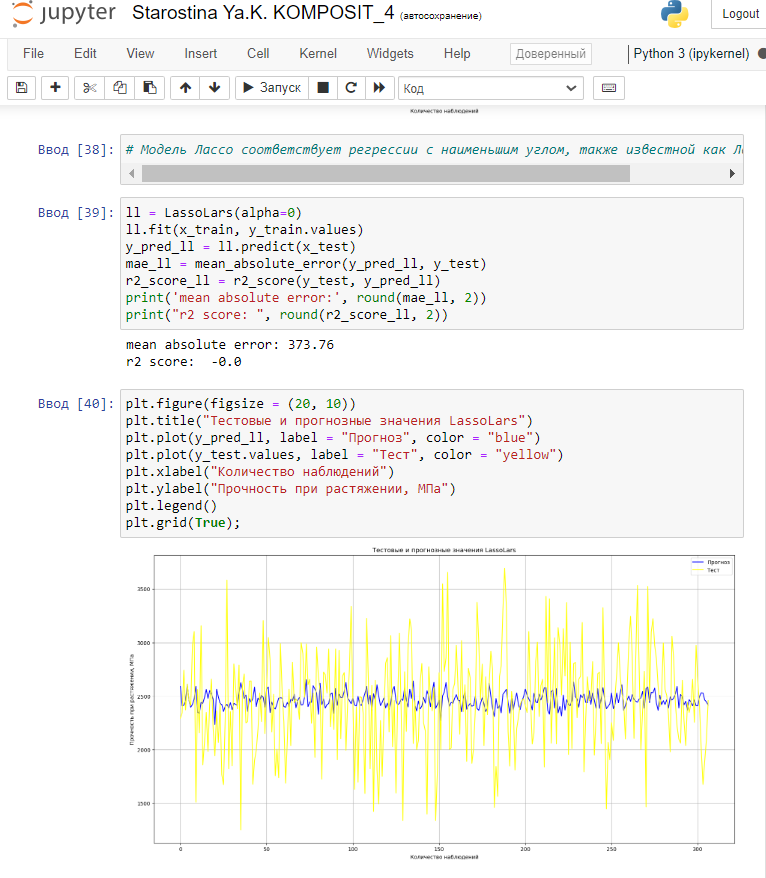
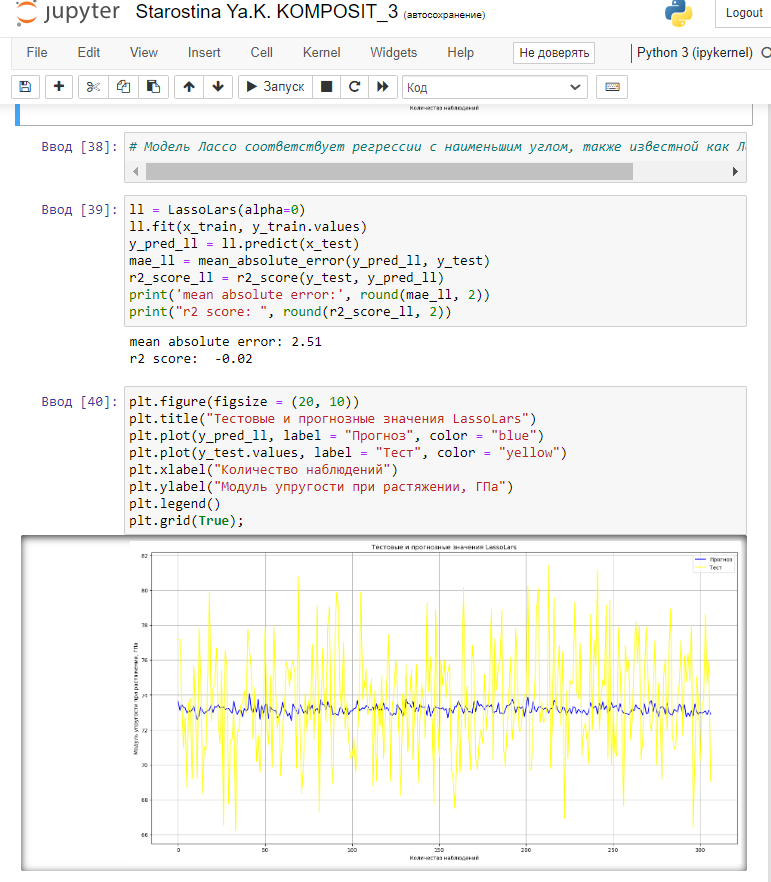


Рисунок 12 – Результаты применения модели Лассо (регрессия с наименьшим углом) для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

Достоинства метода: легко полностью избавляется от шумов в данных; быстро работает; не очень энергоёмко; способно полностью убрать признак из датасета; доступно обнуляет значения коэффициентов.

Недостатки метода: выбор модели не помогает и обычно вредит; часто страдает качество прогнозирования; выдаёт ложное срабатывание результата; случайным образом выбирает одну из коллинеарных переменных; не оценивает правильность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

* + 1. **Модель регрессия байесовского гребня - Bayesian Ridge**

Методы байесовской регрессии можно использовать для включения параметров регуляризации в процедуру оценки: параметр регуляризации не задается в жестком смысле, а настраивается на имеющиеся данные. Результаты применения метода представлены на рисунке 13. Преимущества байесовской регрессии: он адаптируется к имеющимся данным; его можно использовать для включения параметров регуляризации в процедуру оценки. К недостаткам байесовской регрессии относятся: вывод модели может занять много времени.

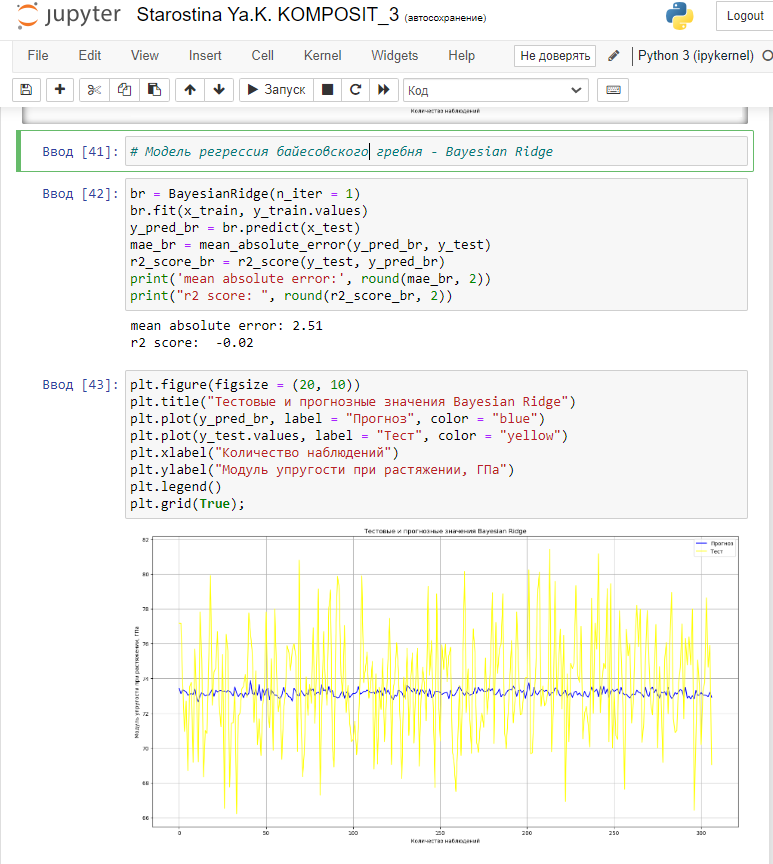
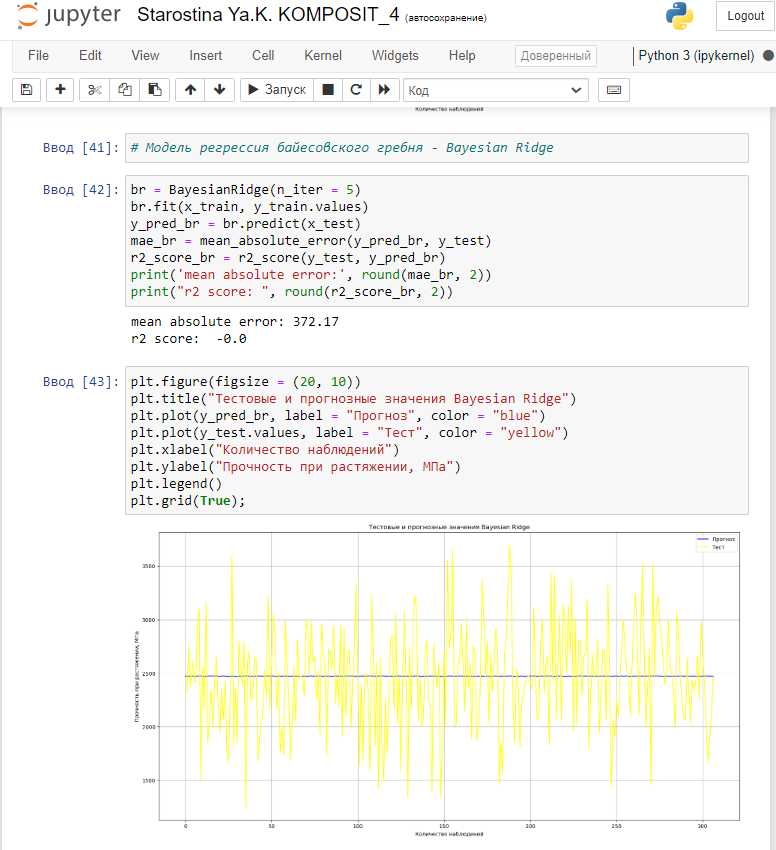
 

Рисунок 13 – Результаты применения модели регрессии байесовского гребня для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

* + 1. **Линейная модель с распределением Пуассона - Poisson Regressor**

Пуассоновская регрессия является представителем большой (и крайне важной) группы регрессионных моделей, которая называется обобщенные линейные модели (GLM), в которую также входит логистическая регрессия. Результаты применения метода представлены на рисунке 14.

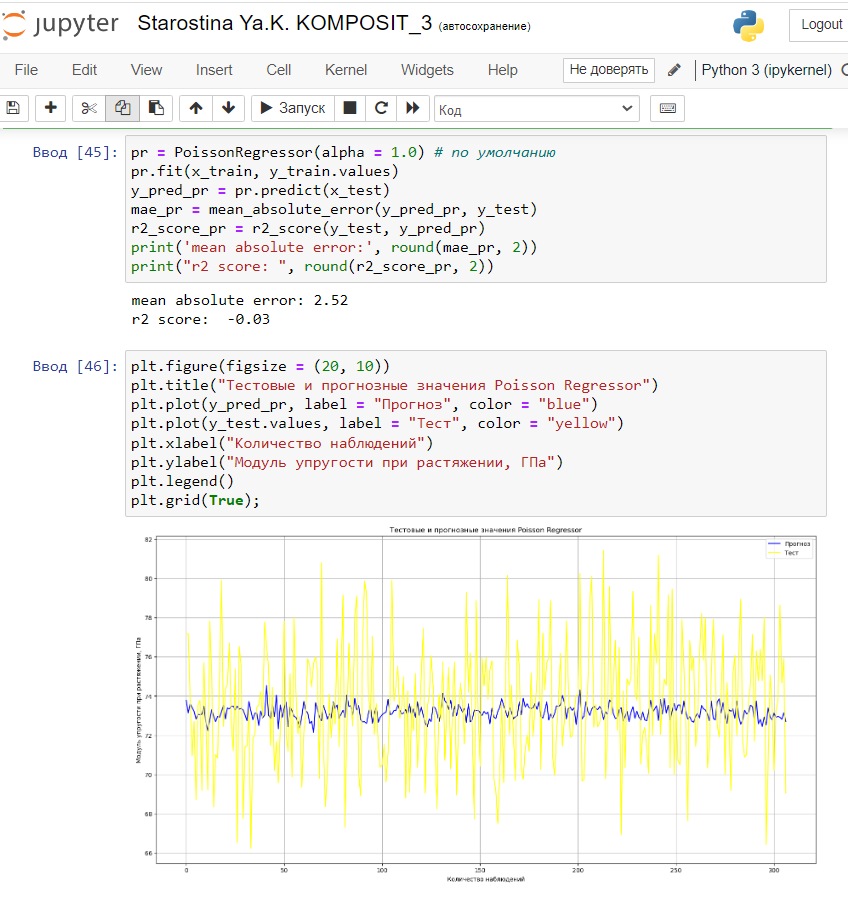
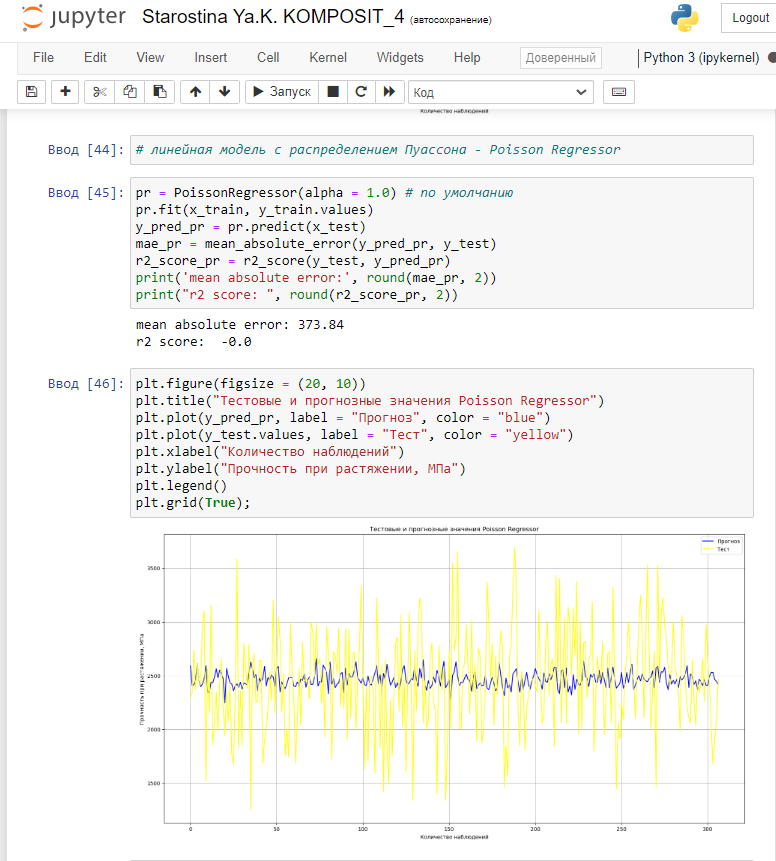
 

Рисунок 14 – Результаты применения модели регрессии с распределением Пуассона для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

Все представители этого семейства имеют следующую структуру:

Случайный компонент – модель для зависимой переменной (в нашем случае для моделирования количества дней в стационаре мы будем использовать пуассоновское распределение, в общем случае – распределение зависимой переменной может быть смоделировано другим представителем экспоненциального семейства).

Систематический (неслучайный) компонент – всегда линейная комбинация предикторов, независимо от их распределения (в нашем случае он только один – САД). Для того, чтобы представить это, попробуйте вспомнить, что находится с правой стороны в простом линейном регрессионном уравнении.

Функция связи – функция, обеспечивающая линейную связь между 1-ым и 2-ым компонентом (в нашем случае это будет натуральный логарифм).

* + 1. **Многослойный перцептрон - Multi-layer Perceptron regressor**

Многослойный персептрон (MLPRegressor) — это алгоритм обучения с учителем, который изучает функцию f(⋅):Rm→Ro обучением на наборе данных, где m — количество измерений для ввода и o- количество размеров для вывода. Это искусственная нейронная сеть, имеющая 3 или более слоёв персептронов. Эти слои - один входной слой, 1 или более скрытых слоёв и один выходной слой персептронов.  Результаты применения метода представлены на рисунке 15.

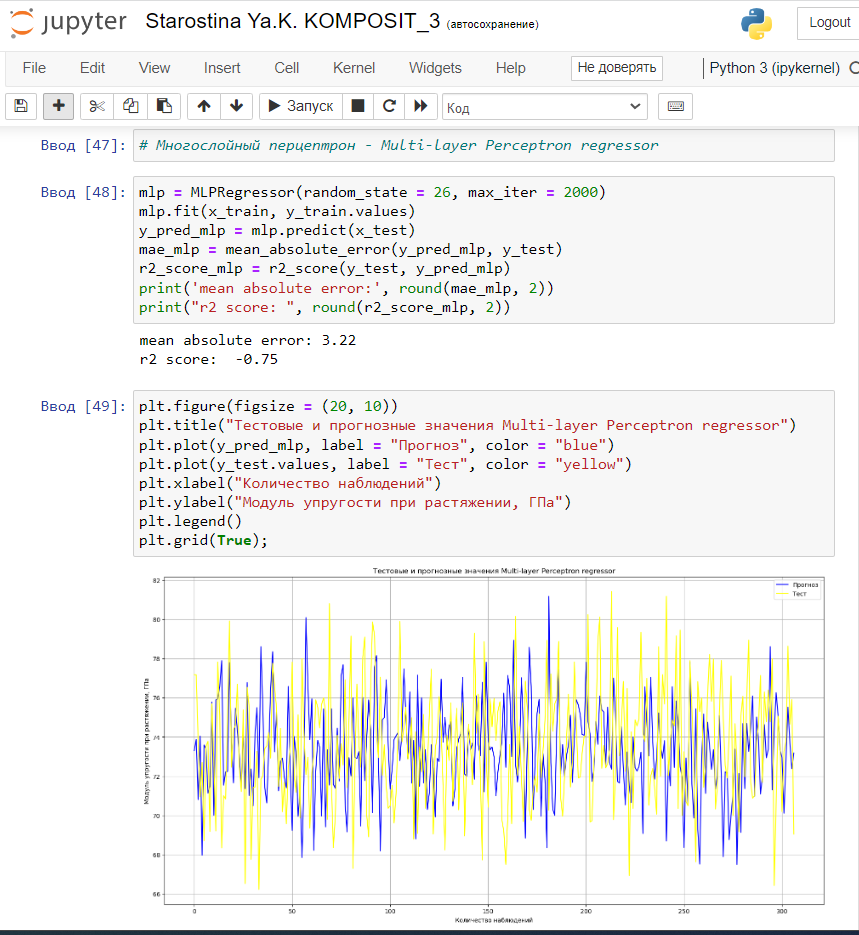
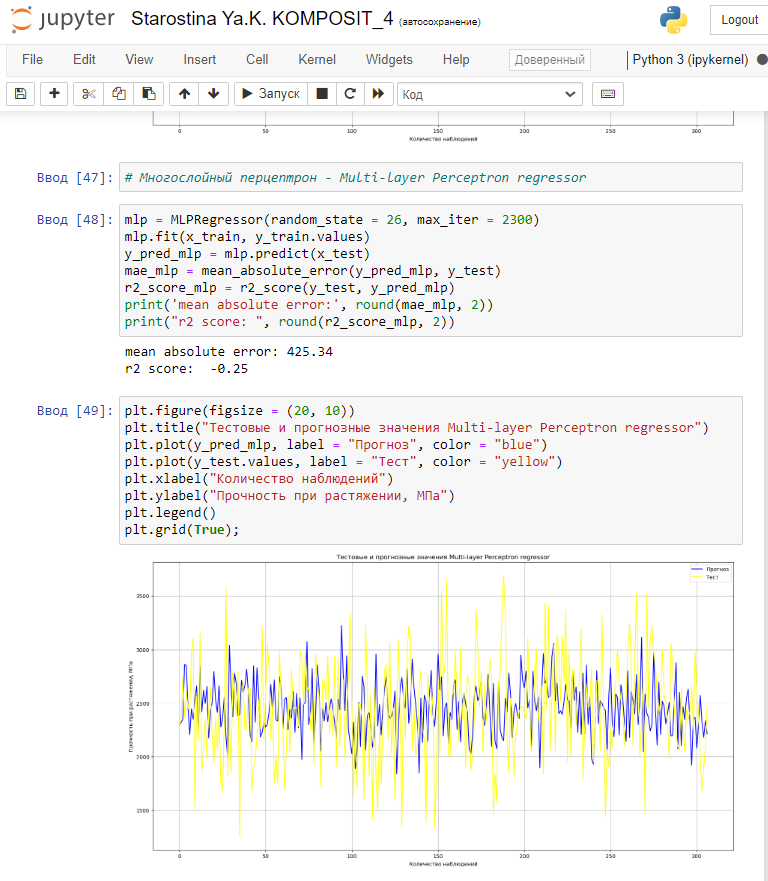
 

Рисунок 15 – Результаты применения многослойный персептрон для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа"

Достоинства метода: построение сложных разделяющих поверхностей; возможность осуществления любого отображения входных векторов в выходные; легко обобщает входные данные; не требует распределения входных векторов; изучает нелинейные модели.

Недостатки метода: имеет невыпуклую функцию потерь; разные инициализации случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

* 1. **Разведочный анализ данных**

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 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.

Цель разведочного анализа - получение первоначальных представлений о характерах распределений переменных исходного набора данных, выявление характера взаимосвязи между переменными с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

В данном разделе приводится краткое описание методов разведочного анализа данных, которые используются для первоначального анализа.

Описательная статистика, вызываемая строкой кода df.describe(), содержит по каждому столбцу (по каждой переменной): count - количество значений mean - среднее значение std - стандартное отклонение min - минимум 25% - верхнее значение первого квартиля 50% - медиана 75% - верхнее значение третьего квартиля max - максимум

Для выявления взаимосвязей параметров в данном по заданию наборе данных были применены следующие коэффициенты корреляции:

* Коэффициент ранговой корреляции Кендалла, обычно называемый коэффициентом Кендалла τ (после греческой буквы τ), представляет собой статистический показатель, используемый для измерения порядковой связи между двумя измеряемыми величинами. Это мера ранговой корреляции: сходство упорядочения данных при ранжировании по каждой из величин. Интуитивно понятно, что корреляция Кендалла между двумя переменными будет высокой, когда наблюдения имеют одинаковый (или идентичный для корреляции 1) ранг (т.е. метку относительного положения наблюдений внутри переменной: 1-й, 2-й, 3-й и т.д.) Между двумя переменными, и низкой, когда наблюдения имеют разный (или полностью отличается для корреляции -1) ранга между двумя переменными.
* Коэффициент корреляции Пирсона, коэффициент корреляции произведение-момент (PPMCC), двумерная корреляция между двумя наборами данных. Это отношение между ковариацией двух переменных и произведением их стандартных отклонений; таким образом, это, по сути, нормализованное измерение ковариации, так что результат всегда имеет значение от -1 до 1. Как и в случае с самой ковариацией, мера может отражать только линейную корреляцию переменных и игнорирует многие другие типы взаимосвязей или корреляций. Форма определения включает в себя "момент произведения", то есть среднее значение (первый момент относительно начала координат) произведения случайных величин с поправкой на среднее значение; отсюда и модификатор product-moment в названии.
* Коэффициент корреляции Спирмена (Spearman rank correlation coefficient) — мера линейной связи между случайными величинами. Корреляция Спирмена является ранговой, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения. Коэффициент корреляции Кенделла и коэффициент корреляции Спирмена выражаются через ранги, но в случае коэффициент корреляции Спирмена инверсиям придаются дополнительные веса (j-i), таким образом коэффициент корреляции Спирмена сильнее реагирует на несогласие ранжировок, чем Коэффициент корреляции Кенделла.

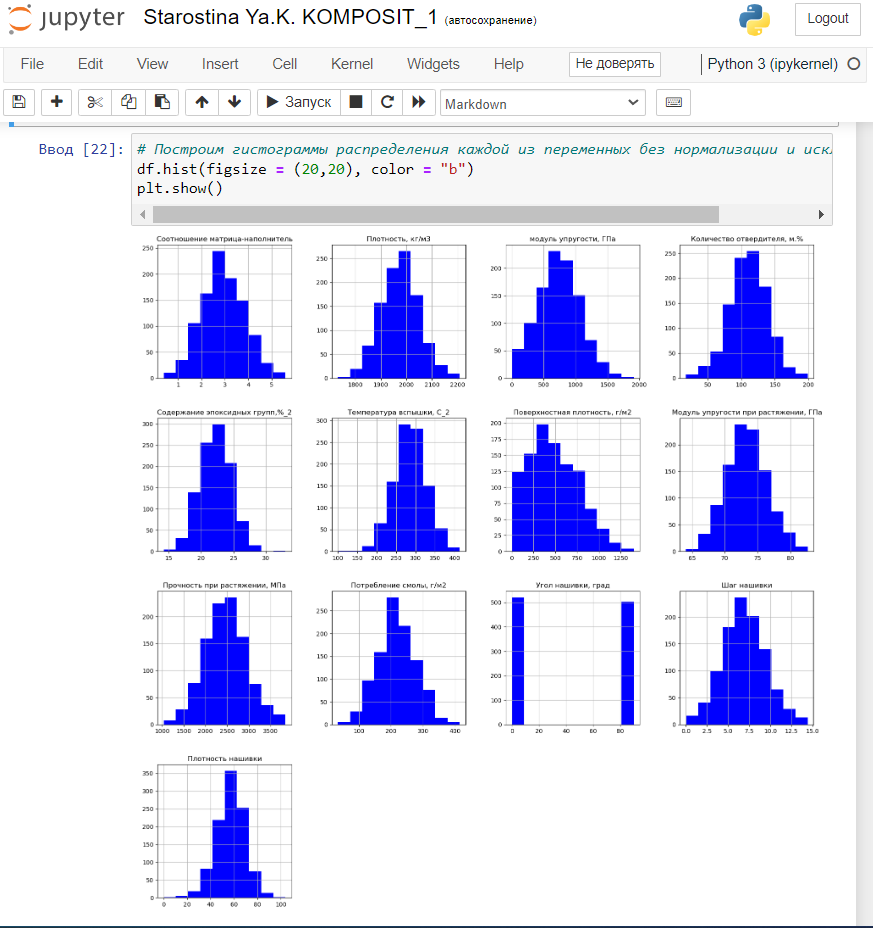


Рисунок 16 – Гистограммы распределения каждой из переменных без нормализации.

Показатели описательной статистики и визуализация гистограмм и/или диаграмм размаха («ящик с усами») позволяют получить наглядное представление о характерах распределений переменных. Такое частотное распределение показывает, какие именно конкретные значения или диапазоны значений исследуемой переменной встречаются наиболее часто, насколько различаются эти значения, расположено ли большинство наблюдений около среднего значения, является распределение симметричным или асимметричным, многомодальным (т.е. имеет две или более вершины) или одномодальным и т.д. По форме распределения можно судить о природе исследуемой переменной

В результате проведённого разведочного анализа данных была выявлена очень слабая корреляция между переменными.

Данные не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рассеяния.

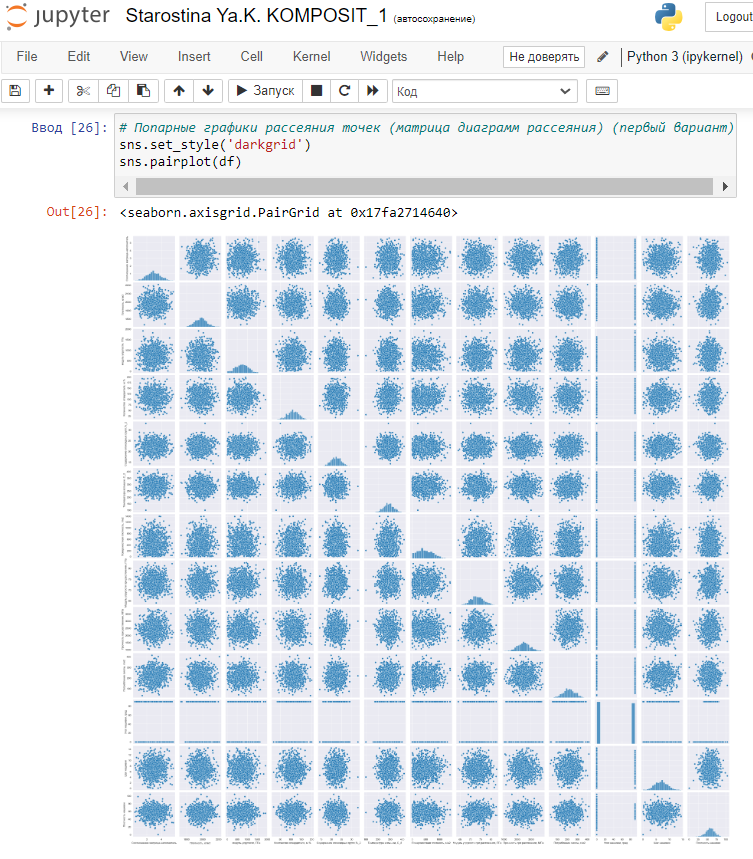
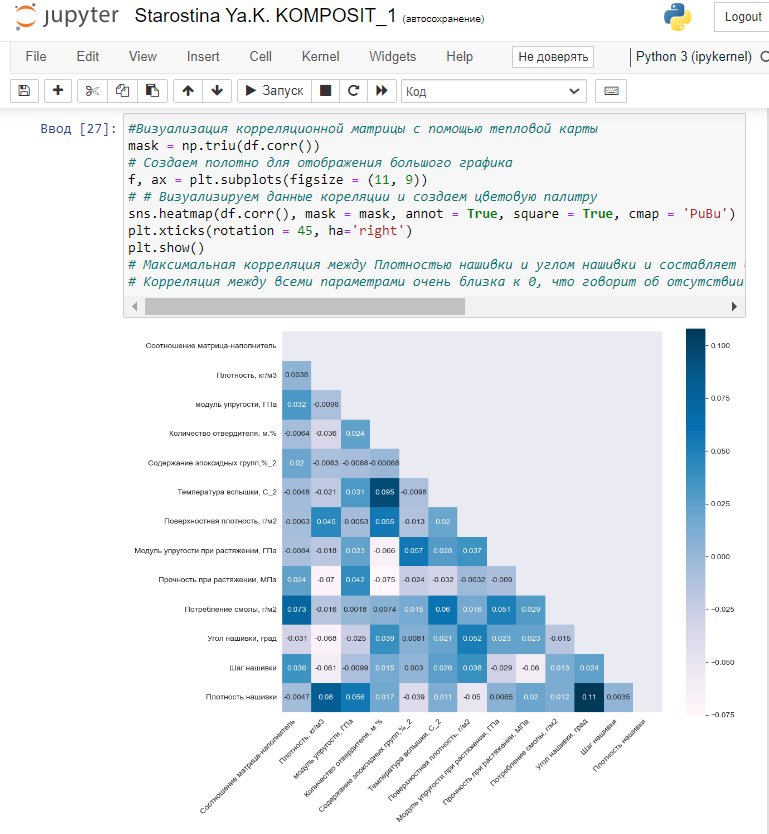
 

Рисунок 17 – Графики диаграмм рассеяния тепловая карта с матрицей корреляции

В соответствии с теорией композитных материалов на качество материла влияет температура вспышки и количество отвердителя из-за взаимодействия отвердителя с матрицей и наполнителем под влиянием температуры. Угол нашивки и плотность нашивки, несомненно, оказывают влияние на свойства материала, что подтверждается максимальным значением 0.11. Но в целом корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

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

Алгоритмы машинного обучения, как правило, работают лучше или сходятся быстрее, когда различные функции (переменные) имеют меньший масштаб. Поэтому перед обучением на них моделей машинного обучения данные обычно нормализуются. Нормализация также делает процесс обучения менее чувствительным к масштабу функций. Это приводит к улучшению коэффициентов после тренировки. Этот процесс повышения пригодности функций для обучения путем изменения масштаба называется масштабированием функций. Мы вычитаем минимальное значение из каждой записи, а затем делим результат на диапазон. Где диапазон - это разница между максимальным значением и минимальным значением.

У нас в основном количественные признаки, поэтому можно применить нормализацию (приведение в диапазон от 0 до 1) или стандартизацию (приведение к матожиданию 0, стандартному отклонению 1). Т.к. это в том числе учебная работа, то используем и нормализацию, и стандартизацию.

MinMaxScaler масштабирует и преобразует каждую характеристику индивидуально, чтобы она находилась в заданном диапазоне обучающей выборки, например, между нулем и единицей. Это преобразование часто используется в качестве альтернативы масштабированию единичной дисперсии с нулевым средним значением.

RobustScaler удаляет медиану и масштабирует данные в соответствии с диапазоном квантилей (по умолчанию IQR: межквартильный диапазон). IQR представляет собой диапазон между 1-м квартилем (25-й квантиль) и 3-м квартилем (75-й квантиль). Центрирование и масштабирование выполняются независимо для каждой функции путем вычисления соответствующей статистики по выборкам в обучающем наборе. Медиана и межквартильный диапазон затем сохраняются для использования в более поздних данных с использованием transform метода.

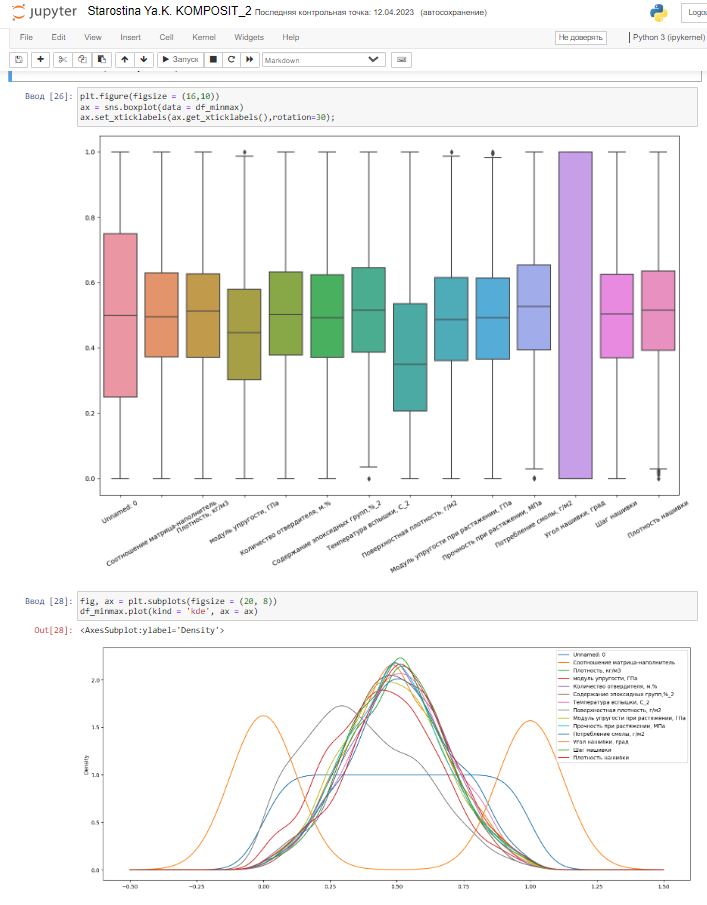
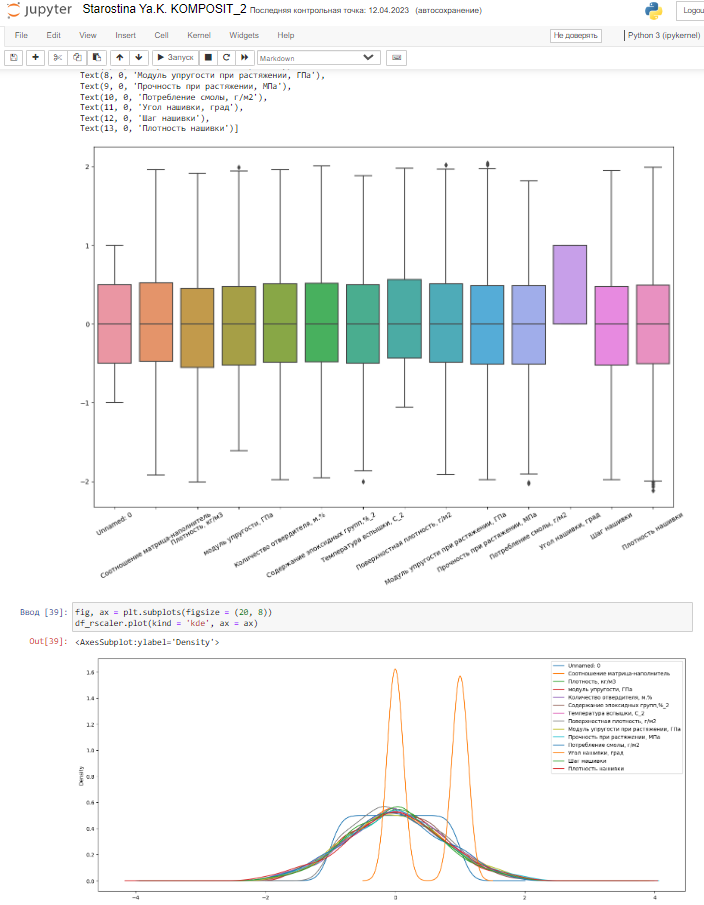
 

Рисунок 18 – Графики и диаграммы размаха при применении MinMaxScaler и RobustScaler соответственно.

* 1. **Разработка и обучение модели**

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно. Для решения применим все методы, описанные выше.

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 70 на 30%); 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.

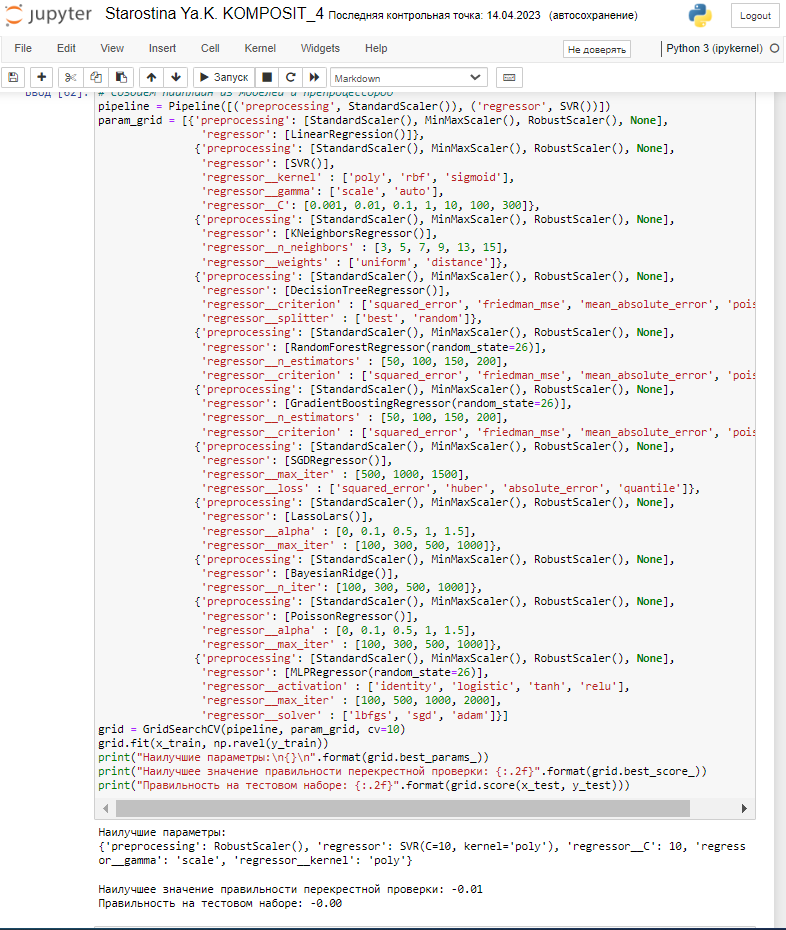
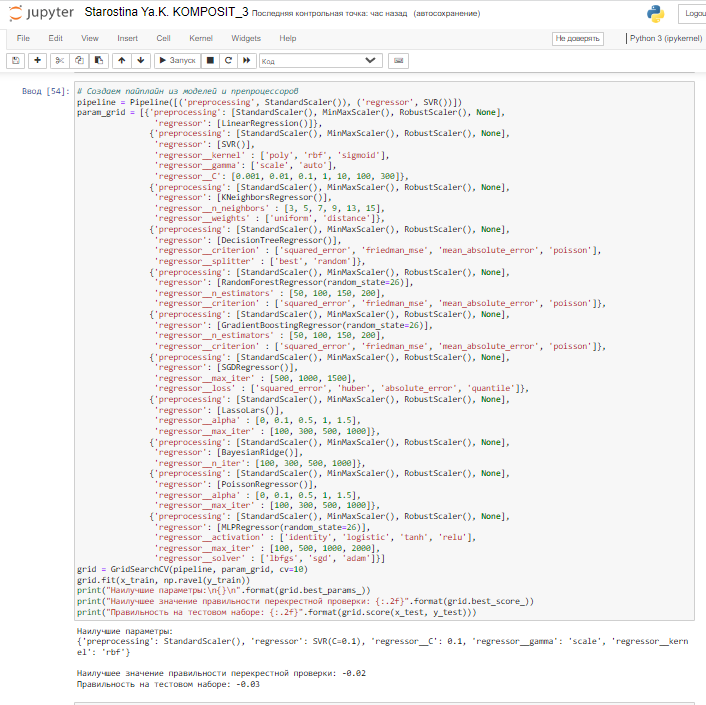


Рисунок 19 – Поиск гиперпараметров для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа" соответственно

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.

Модель после настройки гиперпараметров показала результат немного лучше. Однако, ниже, чем базовая модель. Прочность при растяжении и модуль упругости не имеет линейной зависимости. Все использованные модели не справились с задачей. Результат неудовлетворительный. Свойства композитных материалов в первую очередь зависят от используемых материалов.

* 1. **Тестирование модели**

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя абсолютная ошибка (MАЕ) и коэффициент детерминации (R2). Результат неудовлетворительный.

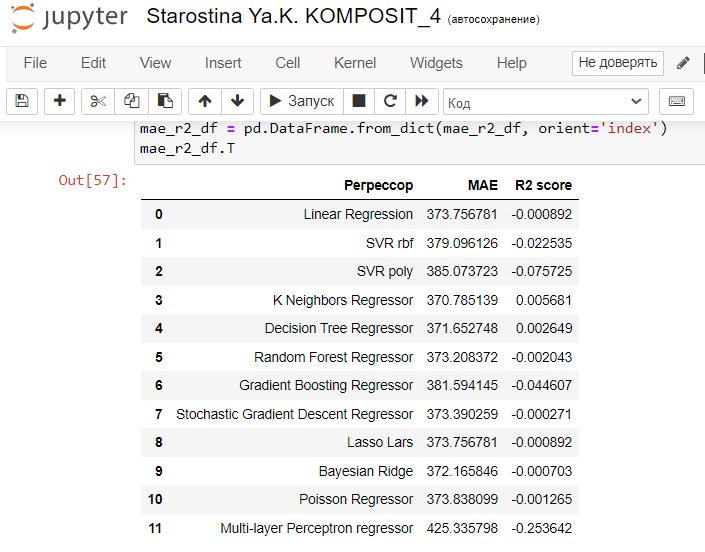
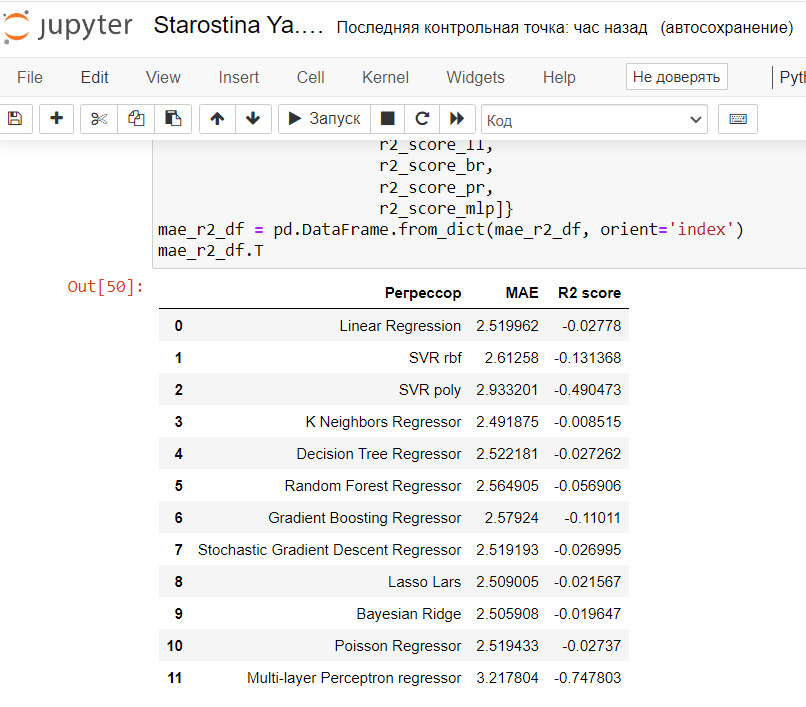


Рисунок 20 - Результаты оценки точности по MАЕ и R2 для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа" соответственно

Хотя в целом при таких результатах можно применять среднее значение переменной в качестве прогнозного.

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

Обучение нейронной сети — это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Построение нейронной сеть было начато с определения лучших параметров при помощи обёртки для API Scikit-Learn.

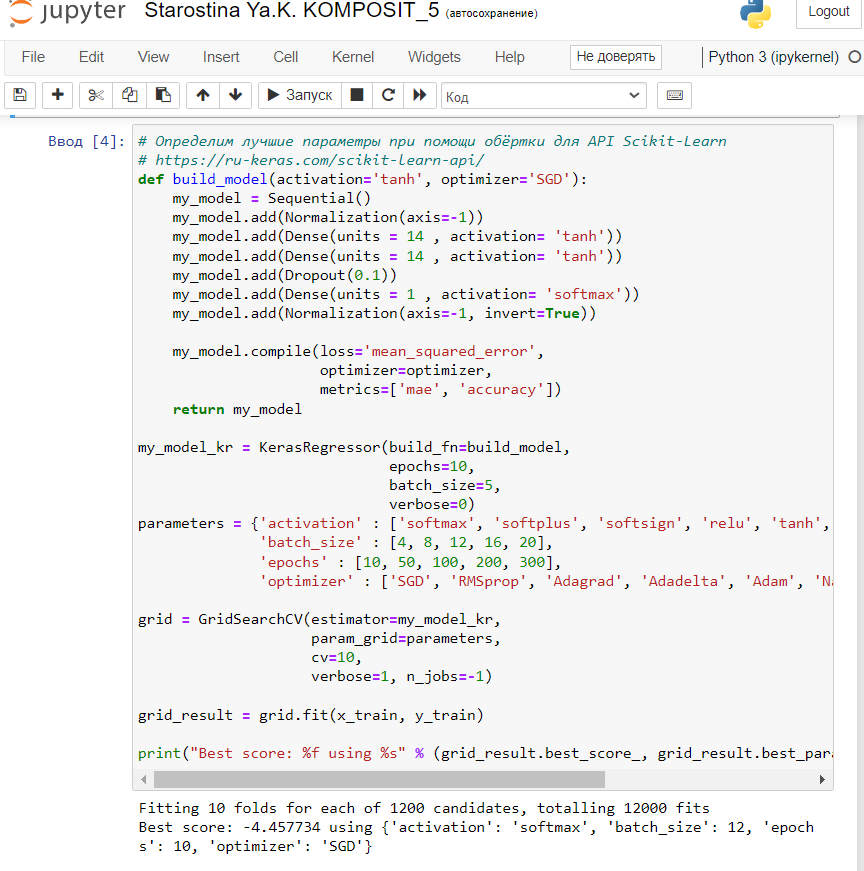


Рисунок 1 – Поиск лучших параметров для создания нейронной сети

После определения параметров построим окончательную нейросеть.

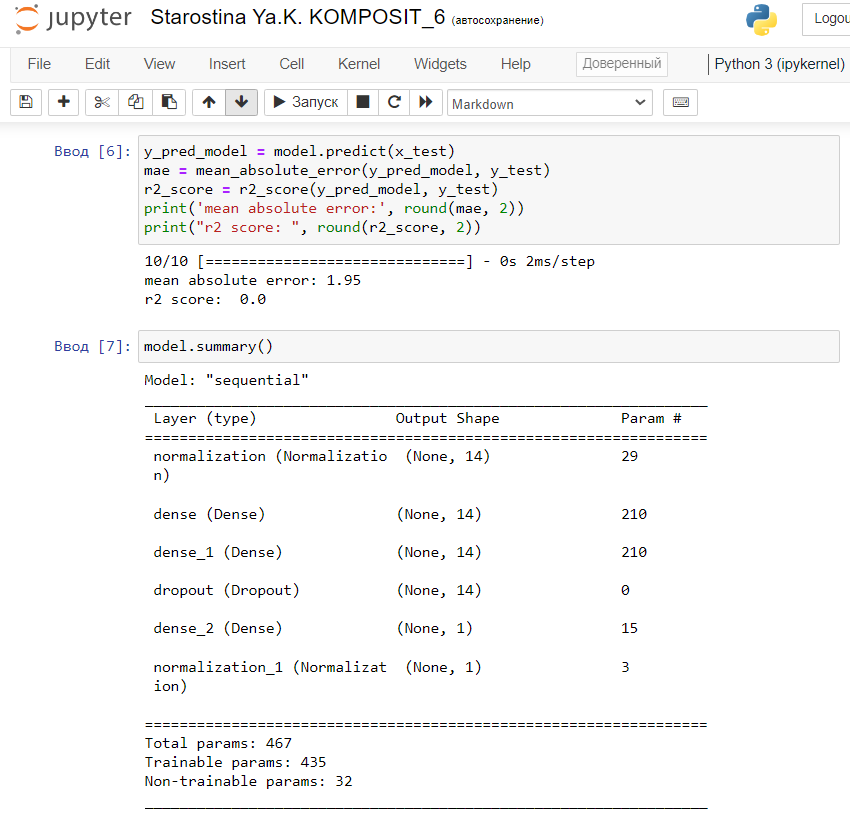
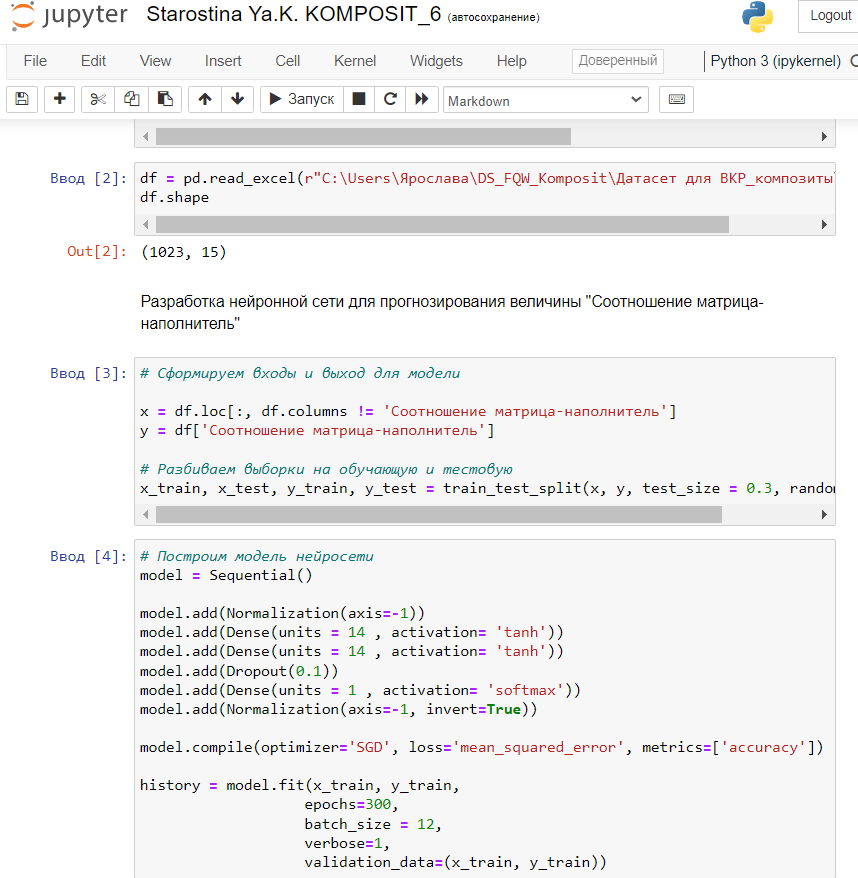


Рисунок 2 - Построение нейронной сети после определения лучших параметров

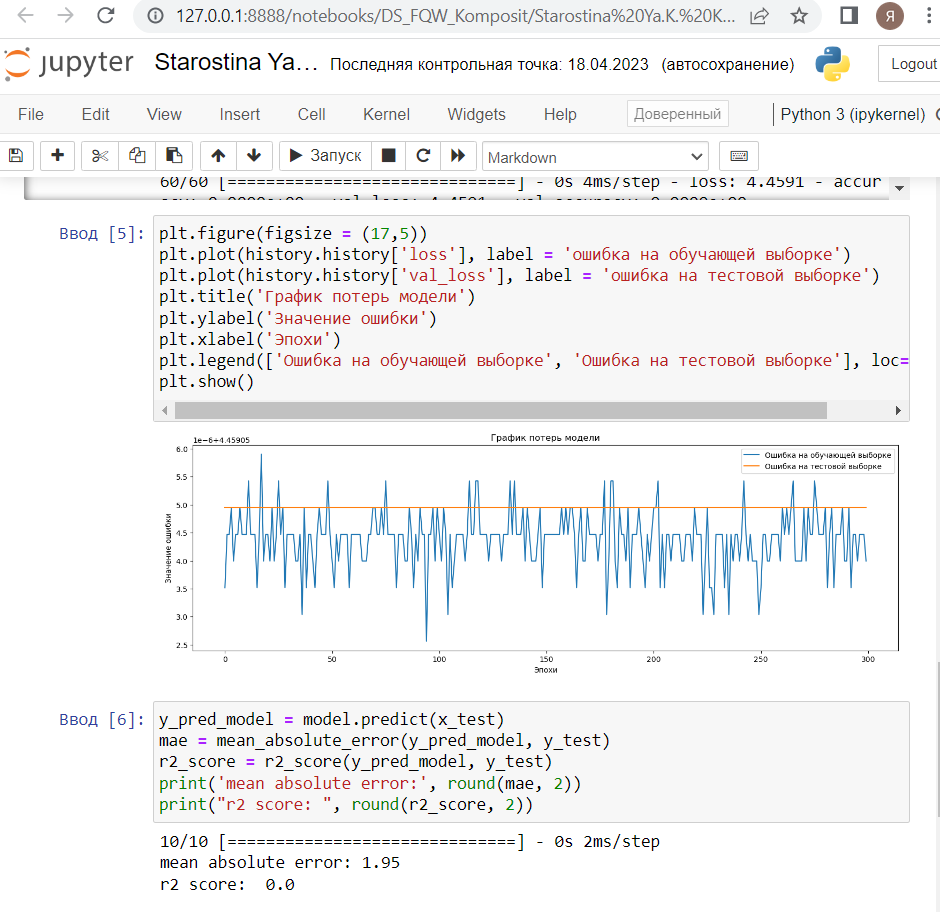


Рисунок 3 - График потерь модели

В результате можно сделать вывод о том, что нейронная сеть обучилась выдавать среднее значение соотношения «матрица-наполнитель».

* 1. **Разработка приложения**

Данное приложение — это основной файл Flask, папка templates, с шаблоном html - страницы, папка s\_model c сохранённой моделью для данных. Часть кода приложения представлена на рисунке 24.

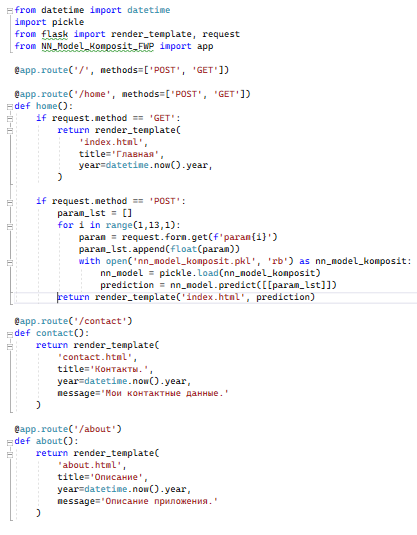


Рисунок 24 - Часть кода приложения

При запуске приложения, пользователь переходит на: <http://127.0.0.1:5555/>.

В открывшемся окне пользователю необходимо ввести в соответствующие ячейки требуемые значения и нажать на кнопку «Отправить».

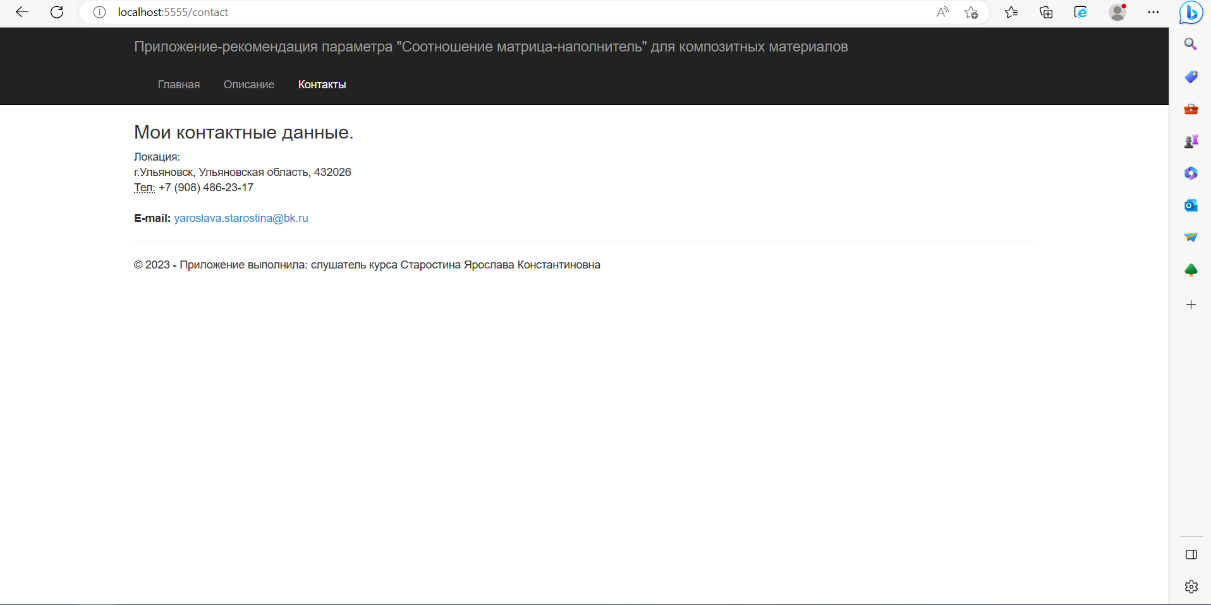
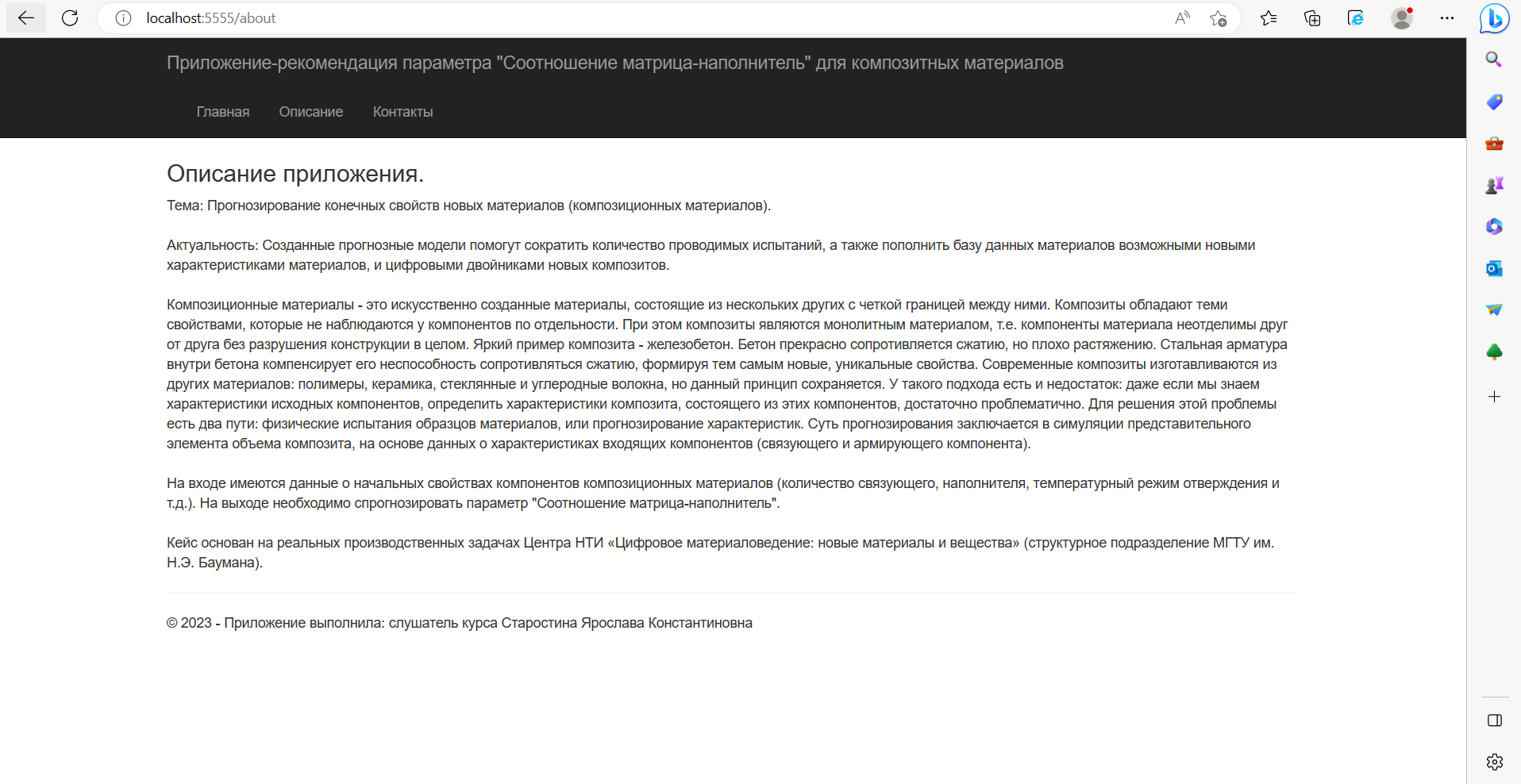
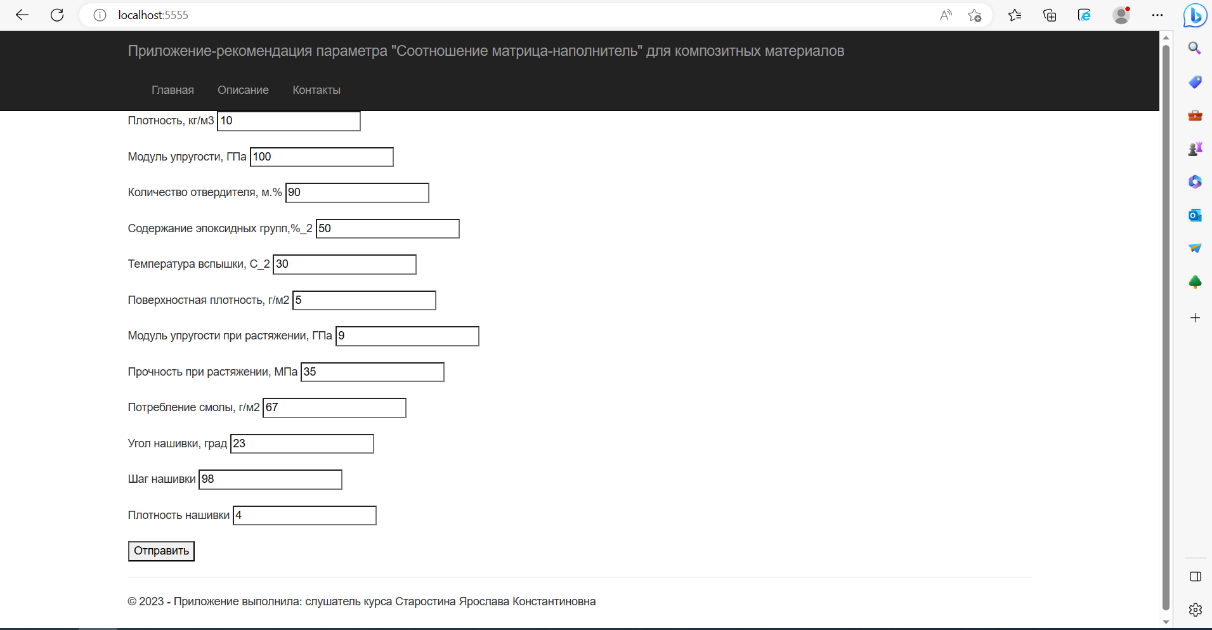


Рисунок 25- Скриншоты веб-приложения

На выходе пользователь получает результат прогноза для значения параметра «Соотношение «матрица – наполнитель»».

* 1. **Создание удалённого репозитория и загрузка**

Репозиторий был создан на github.com по адресу: <https://github.com/YaroSlavaStar/DS_FQW_Komposit>

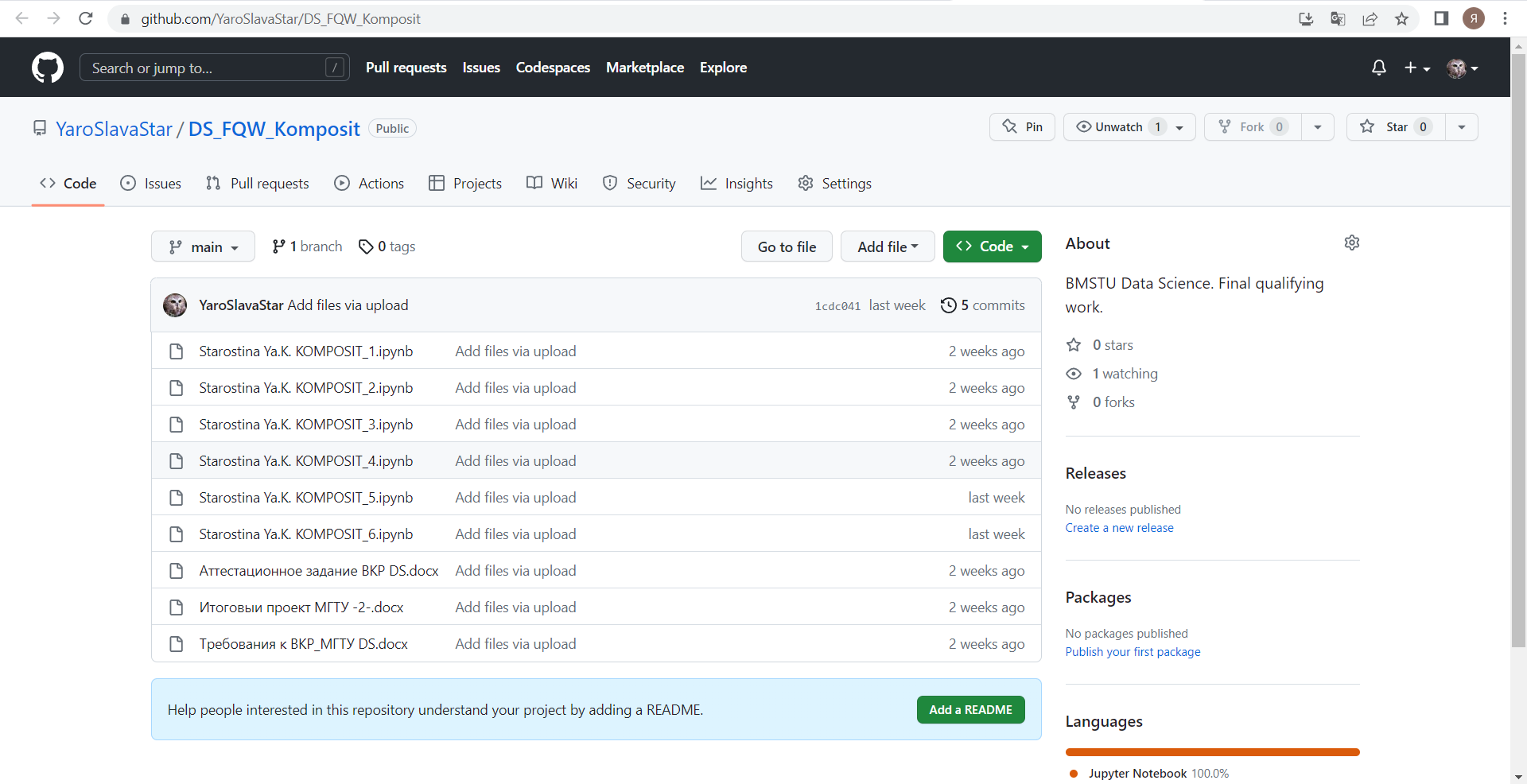


Рисунок 26 – Скрин страницы на github.com

**ЗАКЛЮЧЕНИЕ**

Результаты, полученные при выполнении выпускной квалификационной работы, позволяют сделать следующие выводы:

* распределение полученных данных в объединённом датасете близко к нормальному;
* коэффициенты корреляции между парами признаков стремятся к нулю;
* применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов (для прогнозирования величины "Модуль упругости при растяжении, ГПа" и "Прочность при растяжении, МПа" лучший метод – Метод опорных векторов);
* разработанная модель нейронной сети, с учётом подборки лучших параметров, не позволила получить достоверных прогнозов.

Полученные результаты указывают на то, что необходимы дополнительные вводные данные, консультации экспертов предметной области, новые исследования.

**СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ**

1. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
2. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.
3. 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.
4. 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.
5. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.
6. 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.
7. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.
8. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
9. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.