Образовательный центр МГТУ им. Баумана

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

Data Science

“Прогнозирование конечных свойств

новых (композиционных) материалов”

Слушатель: Закирова Ю.Л.

Москва, 2023

Содержание

[Введение 3](#_Toc127559069)

[1 Аналитическая часть 5](#_Toc127559070)

[1.1 Постановка задачи 5](#_Toc127559071)

[1.2 Описание используемых методов 9](#_Toc127559072)

[1.2.1 Линейная регрессия 10](#_Toc127559073)

[1.2.2 Метод опорных векторов 11](#_Toc127559074)

[1.2.3 Метод случайного леса 12](#_Toc127559075)

[2.3 Разведочный анализ данных 13](#_Toc127559076)

[2 Практическая часть 15](#_Toc127559077)

[2.1 Предобработка данных 15](#_Toc127559078)

[2.2 Разработка, обучение и тестирование модели 17](#_Toc127559079)

[2.3 Построение нейронной сети для рекомендации соотношения матрица-наполнитель 20](#_Toc127559080)

[2.4 Разработка приложения 21](#_Toc127559081)

[Заключение 27](#_Toc127559082)

[Библиографический список 29](#_Toc127559083)

Приложение ………………………………………………………………………..31

# **Введение**

Тема данной работы - прогнозирование конечных свойств композиционных материалов.

Композиционные материалы – это материалы, состоящие из двух или более компонентов, каждый из которых имеет свои уникальные свойства. Такие материалы могут быть использованы для создания легких и прочных автомобильных кузовов, а также для производства легких и прочных корпусов для бортовых компьютеров.

Композиционные материалы могут быть изготовлены из различных типов материалов, включая металлы, пластмассы, древесину, карбон и стекло. Каждый из этих материалов имеет свои уникальные свойства, которые могут быть использованы для создания более прочных и производительных продуктов. Например, металлы могут предоставить дополнительную прочность и производительность, а пластмассы могут быть использованы для создания легких и прочных корпусов.

Кроме того, композиционные материалы могут быть использованы для создания сложных и прочных конструкций, которые могут выдерживать высокие нагрузки и противостоять воздействиям окружающей среды. Такие конструкции могут быть использованы для строительства мостов, зданий и даже космических аппаратов.

Учитывая такое широкое распространение и высокую потребность в новых материалах, тема данной работы является очень актуальной.

Методы машинного обучения представляют собой набор алгоритмов и стратегий, которые используются для автоматического извлечения знаний из данных. Они помогают компьютеру автономно изучать и понимать данные, а так же стоить прогнозы для будущих результатов.

Одним из самых популярных методов машинного обучения является глубокое обучение. Этот метод использует нейронные сети для обучения алгоритмов и позволяет им изучать данные более глубоко и точно.

В целом, методы машинного обучения являются мощными инструментами для анализа данных и предсказания будущих результатов. Они помогают автоматизировать процесс извлечения знаний из данных и позволяют компьютеру быстро и эффективно решать различные практические задачи.

Стоимость производства композитного материала достаточно высока. Без изучения характеристик составляющих компонентов невозможно предсказать свойства композита. Поэтому для достижения желаемых свойств требуется проведение большого количества испытаний различных комбинаций. Система поддержки производственных решений, основанная на принципах машинного обучения, могла бы существенно сократить время и затраты на создание определенного материала.

# **1 Аналитическая часть**

## **1.1 Постановка задачи**

В ходе данной работы была выстроена модель машинного обучения, способная прогнозировать неизвестные свойства композитного материала, в зависимости от исходных входных данных. Исходными данными являются свойства, в зависимости от которых необходимо определить недостающие значения характеристики материала. Так как композиционный материал это неоднородный материал, состоящий из нескольких компонентов, в котором отображается граница раздела входящих компонентов. Анализ предметной области позволил выделить свойства композитных материалов их численное значения для формирования двух таблиц данных X\_nup.xlsx и X\_bp.xlsx.

В первом файле хранится информация о составляющей из углепластика. На рисунке 1 представлено 10 первых строк и 10 последних файла X\_nup.xlsx.

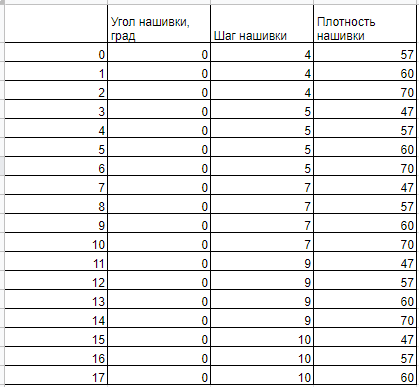
Файл X\_nup содержит 1014 строк, 3 признака и индекс.

Во втором файле X\_bp.xlsx хранятся данные о составляющей из углепластика. На рисунке 2 представлено 10 первых строк и 10 последних файла X\_bp.xlsx.

Файл X\_bp.xlsx содержит 1023 строк, 10 признаков и индекс.

Для удобства два файла были объединены в один DataFrame, в соответствии с рисунком 3.

В отличие от обычной таблицы, DataFrame может содержать данные из различных типов, в том числе целые числа, строки, логические значения и даты. Это делает ее идеальным инструментом для анализа данных. Она предоставляет мощные возможности для группировки и анализа данных, что позволяет быстро и эффективно принимать решения.



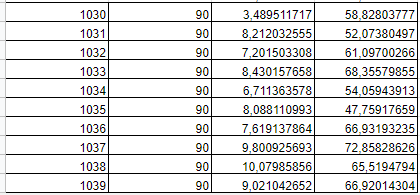
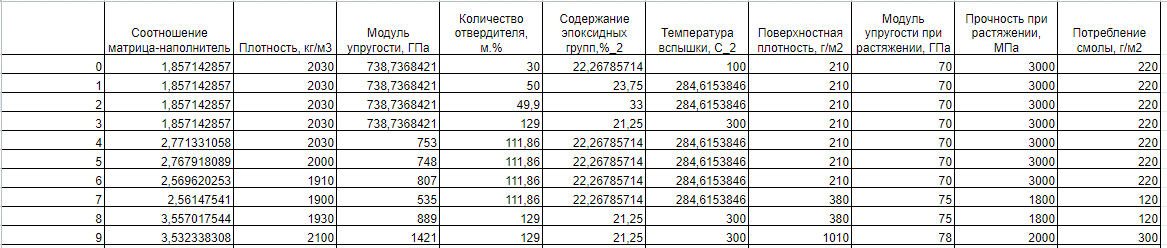


Рисунок 1- Часть данных из файла X\_nup.xlsx



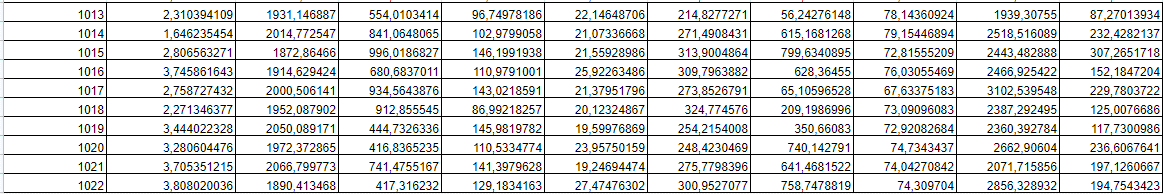


Рисунок 2- Часть данных из файла X\_bp.xlsx

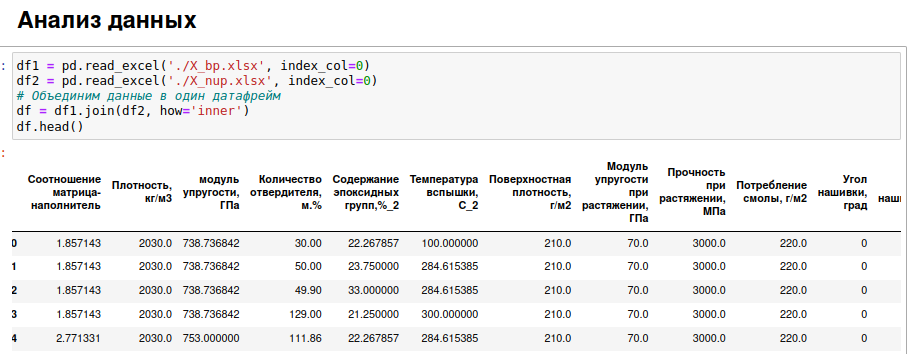


Рисунок 3- Данные, объединенные в DataFrame

В ходе объединения данных часть строк данных о составляющей из углепластика была отброшена. Следовательно, конечный датасет состоит из 1023 строк и 13 признаков. Благодаря предварительному анализу, было выявлено, что в данных не имеется пропусков – значит, все строки используются для дальнейшей работы. Все признаки имеют числовой тип данных.

Датасеты являются одним из основных инструментов для анализа данных. Они представляют собой наборы данных, которые могут использоваться для построения моделей и проведения анализа. Датасеты могут содержать данные любого типа, включая текстовые, цифровые, графические и аудиовизуальные данные.

При детальном исследовании данных были обнаружены выброы в соответствии с рисунком 4.

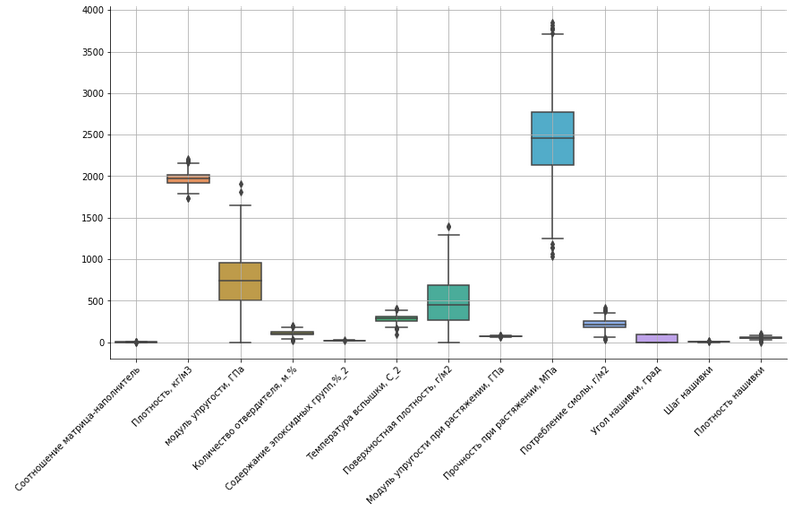


Рисунок 4- Диаграмма размаха (“ящик с усами”)

Выбросы в первоначальной тренировочной выборке в соответствии с рисунком 4 приводят к ухудшению обобщающей способности модели машинного обучения, следовательно, необходимо провести фильтрацию с помощью границ Тьюки.

Алгоритм метода Тьюки состоит из некоторых действий. Сперва необходимо рассчитать 1-й и 3-й квартиль (или, что тоже самое, 25-й и 75-й перцентиль) ряда данных, в котором требуется найти выбросы. Затем вычесть первый квартиль из третьего — получится мера распределения данных, называемая межквартильным размахом (МР). МР устойчив к экстремальным значениям распределения (то есть робастный), в отличие от стандартной ошибки, которая более чувствительна к выбросам. А также рассчитать нижнюю и верхнюю внутренние границы, отстоящие на 1,5МР от квартилей. Далее необходимо рассчитать нижнюю и верхнюю внешние границы, отстоящие на 3МР от квартилей. Значение за пределами границ – экстремально, их нужно отбросить.

В результате данной фильтрации количество записей уменьшилось до 932. Обновлённый набор данных не содержит выбросов в соответствии с рисунком 5.

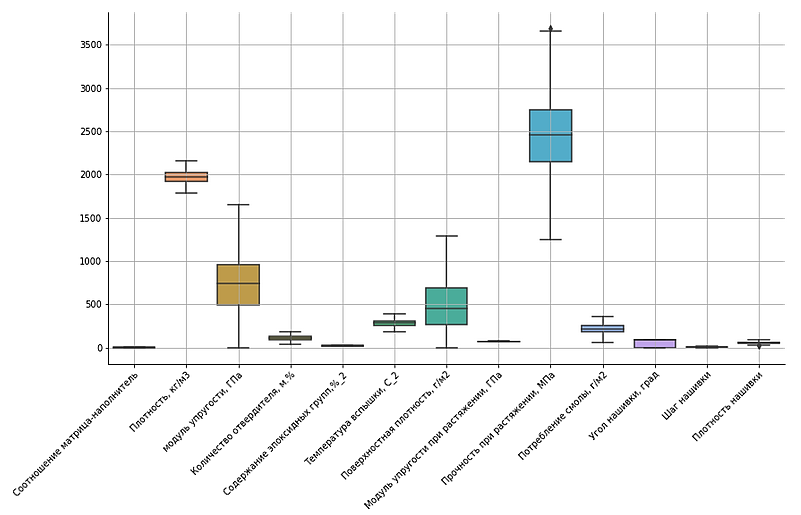


Рисунок 5- результат фильтрации методом Тьюки

Анализ попарной корреляции признаков (в соответствии с рисунком 5) показал отсутствие сильных зависимостей, поэтому все признаки считаются релевантными и не будут отброшены.

## **1.2 Описание используемых методов**

Существует множество различных методов машинного обучения. Они могут быть классифицированы как супервизионные или несупервизионные. Супервизионные методы машинного обучения используют предварительно определенные правила для обработки данных. Например, алгоритм логического решения использует предварительно определенные правила для построения модели. Несупервизионные методы машинного обучения используют алгоритмы, которые могут автоматически извлекать закономерности из данных.

В качестве моделей для машинного обучения в соответствии с поставленной задачей были выбраны такие методы как: линейная регрессия, метод опорных векторов и случайный лес.

### **1.2.1 Линейная регрессия**

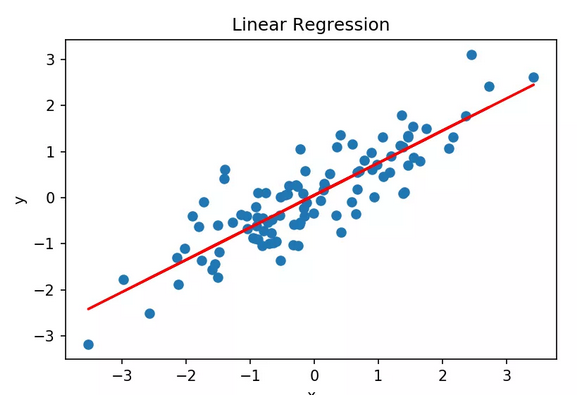


Рисунок 6- Иллюстрация метода линейной регрессии

При рассмотрении зависимости между одной входной и одной выходной переменных используется регрессионная модель. Предельные свойства оценки нелинейных моделей выводятся в результате аппроксимации с помощью линейных моделей. При методе линейной регрессии определяется уравнение регрессии и строится прямая-линия регрессии по формуле (1), в соответствии с рисунком 6.

, (1)

где y - прогнозируемое значение, m - коэффициент при x, x - значение одной переменной, b - свободный член.

Для построения модели линейной регрессии необходимо найти оптимальные значения коэффициентов m и b, для этого используется метод наименьших квадратов. Этот метод позволяет найти наилучшее соответствие между данными и линейной функцией.

Метод линейной регрессии имеет ряд преимуществ, включая простоту использования, быстроту обучения и высокую точность прогнозирования. Однако он также имеет некоторые недостатки, включая неспособность обрабатывать большие и сложные данные и ограниченную применимость к нелинейным задачам.

### **1.2.2 Метод опорных векторов**



Рисунок 7- Иллюстрация метода опорных векторов

Для задач классификации и регрессионного анализа также используется метод опорных векторов (SVM). Метод опорных векторов к тому же известен как метод классификатора с максимальным зазором. Из-за того, что особенность метода является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора.

Данный метод базируется на переводе исходных векторов в пространство более высокой размерности, затем происходит поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы в соответствии с рисунком 7.

Одним из главных преимуществ метода SVM является его способность достичь высокой точности при классификации данных. Он может работать с любым набором данных, независимо от размера и сложности. Также он может быть использован для построения моделей, которые могут предсказывать будущие результаты.

Несмотря на множество преимуществ, метод SVM имеет также некоторые недостатки. Он может быть достаточно сложен для понимания и может требовать больших объемов данных для обучения. Также он может быть медленным при обработке больших объемов данных.

### **1.2.3 Метод случайного леса**

Алгоритм машинного обучения, основанный в использовании ансамбля решающих деревьев в соответствии с рисунком 8 называют методом случайного леса.

Алгоритм содержит в себе идеи метода бэггинга Бреймана и метода случайных пространств. Алгоритм находит свое применение в задачах классификации, кластеризации и регрессии.

Идея метода базируется на использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим. Каждое дерево леса строится независимо друг от друга. Оптимальное количество деревьев подбирается так, чтобы число ошибок классификатора на тестовой выборке было минимальным.

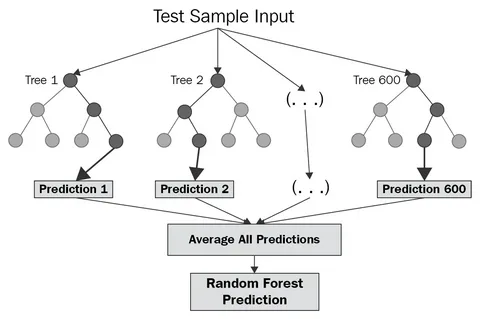


Рисунок 8- Иллюстрация метода случайный лес

Одним из главных преимуществ метода случайного леса является его способность предотвращать переобучение. Это достигается за счет того, что каждое дерево использует случайно выбранные признаки, что позволяет избежать переобучения и предоставить более надежные прогнозы.

В целом, метод случайного леса – это мощный алгоритм машинного обучения, который предоставляет надежные прогнозы и предотвращает переобучение.

## **2.3 Разведочный анализ данных**

Цель разведочного анализа данных — выявить закономерности в данных. Для корректной работы большинства моделей желательна сильная зависимость выходных переменных от входных и отсутствие зависимости между входными переменными.

По матрице корреляции (рисунок 9) мы видим, что все коэффициенты корреляции близки к нулю, что означает отсутствие линейной зависимости между признаками. Матрица корреляции позволяет выявить связь между признаками.

Как описано выше, признаки имеют различных масштаб. Модели машинного обучения показывают лучшую точность, если данных нормированы.

Для масштабирования данных, то есть для их обработки при числовых функциях, используется StandardScaler из пакета sklearn. Средние значения до нормировки и после в соответствии с рисунком 10.

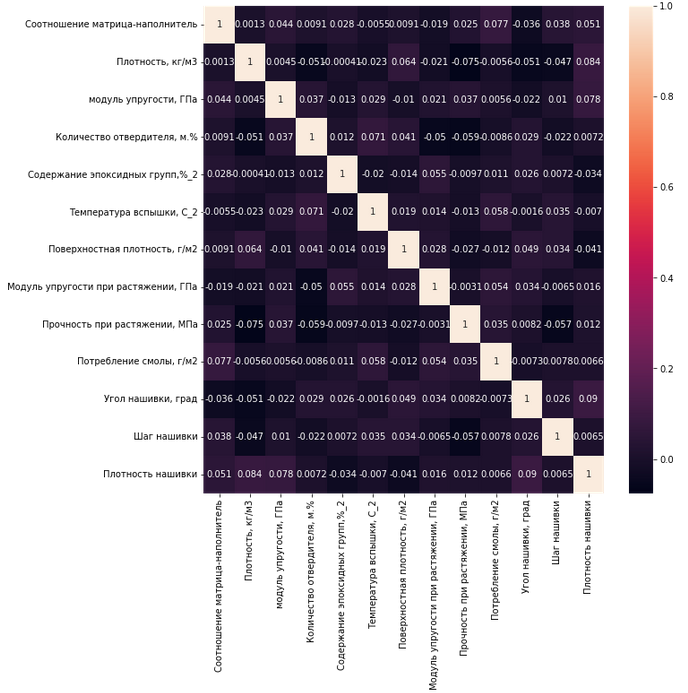


Рисунок 9- Матрица корреляции

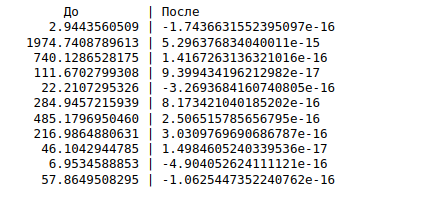


Рисунок 10- Средние значения данных

# **2 Практическая часть**

Материалы загружены на Github ([JuZa2023/VKR: Task (github.com)](https://github.com/JuZa2023/VKR))

## 

## 

Выведем числовые характеристики для каждого признака, такие как:

* количество записей примеров
* среднее значение
* стандартное отклонение
* 25% квантиль
* 50% квантиль
* 75% квантиль
* минимальное значение
* максимальное значение

## 

## 

## 

## 

## 

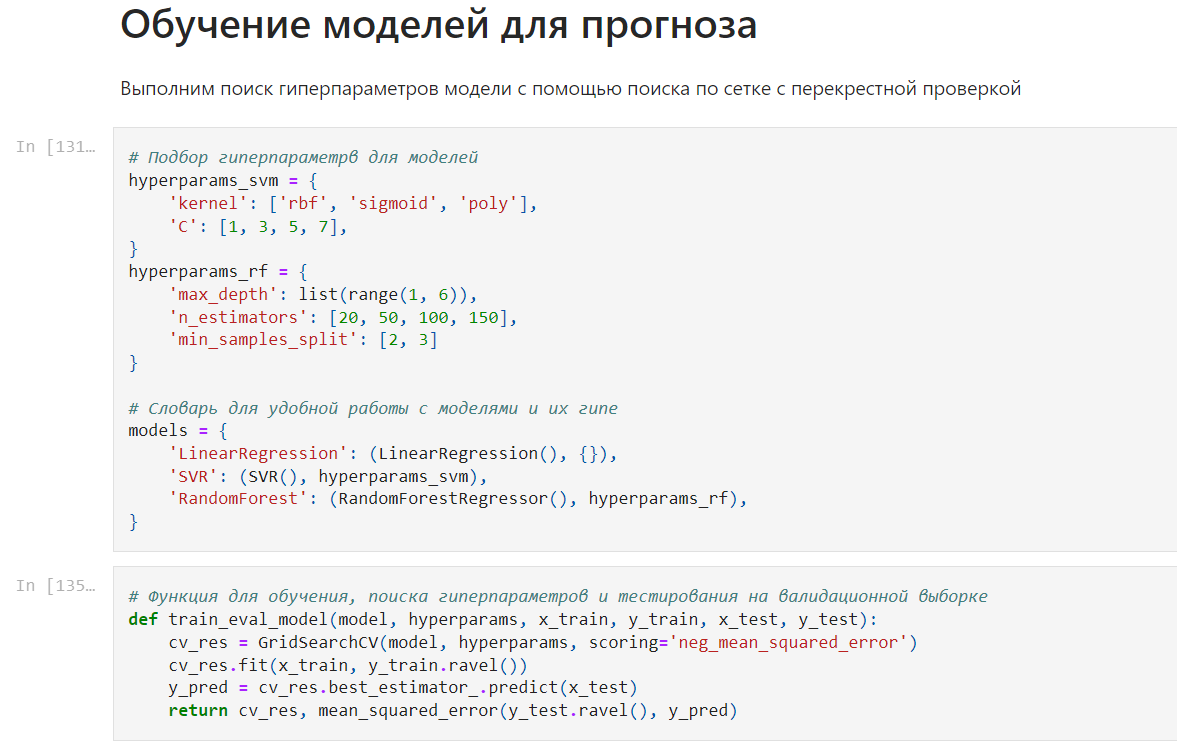
## 

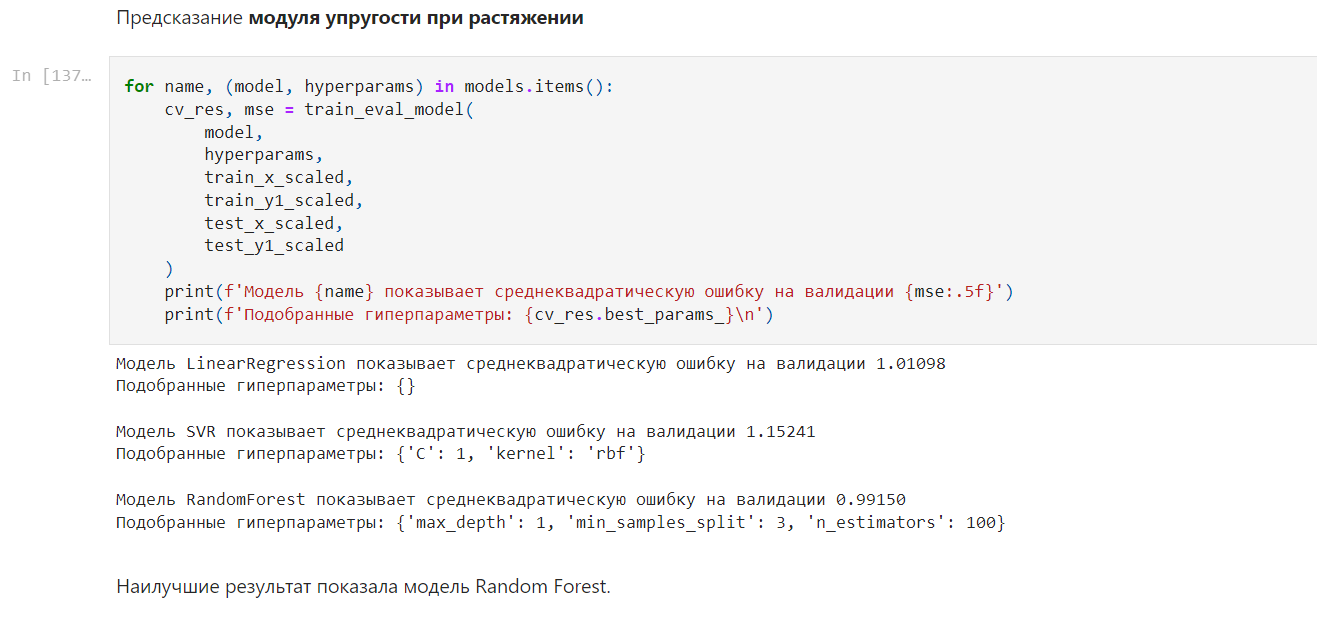
## 

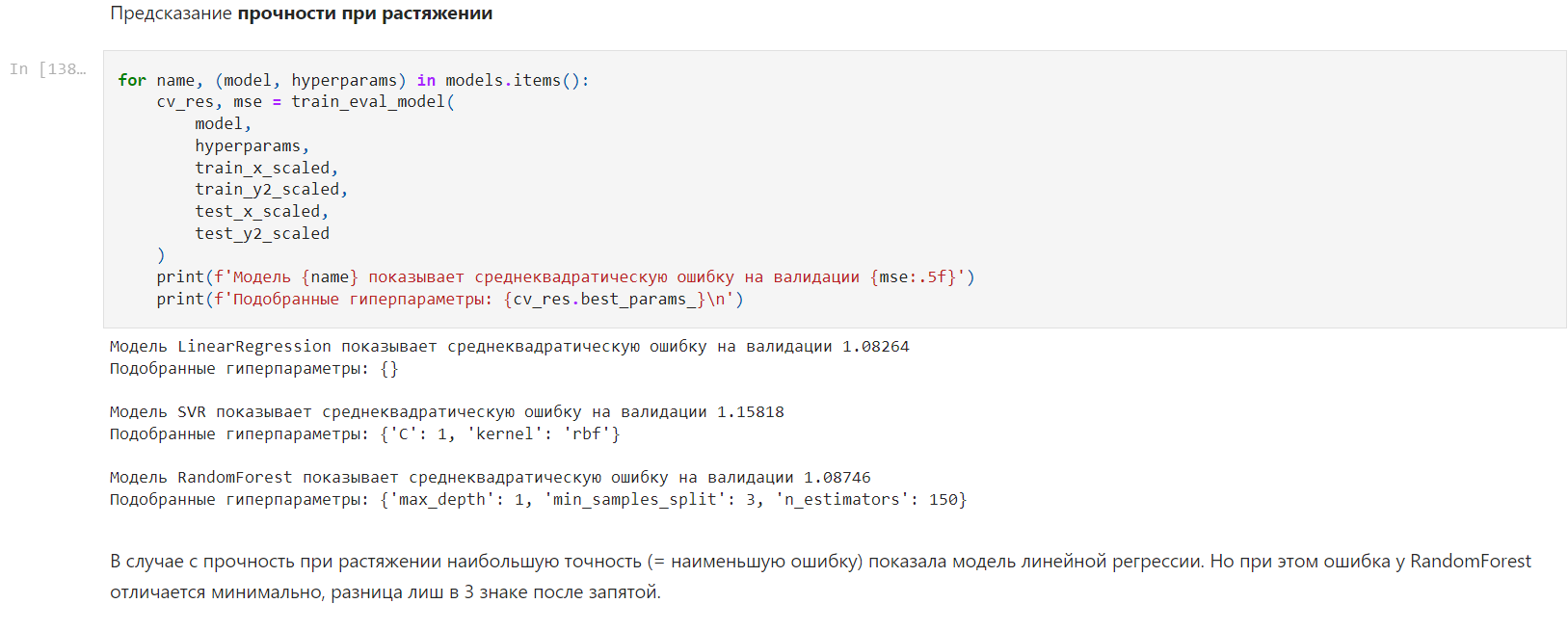
## 

## 

## 







# 

# 

# 

# **Заключение**

Проведенное исследование позволяет рассмотреть большую часть операций и задач, которые приходится выполнять специалисту по работе с данными.

В ходе работы были рассмотрены такие аспекты работы с данными, как:

* изучение теоретических методов анализа данных и машинного обучения;
* изучение основ предметной области, в которой решается задача;
* извлечение и транформация данных;
* проведение разведочного анализа данных статистическими методами;
* DataMining — извлечение признаков из датасета и их анализ;
* разделение имеющихся, в нашем случае размеченных данных на обучающую, валидационную, тестовую выборки;
* выполнение предобработки (препроцессинга) данных для обеспечения корректной работы моделей;
* построение аналитического решения (это включает выбор алгоритма решения и модели, сравнение различных моделей, подбор гиперпараметров модели);
* визуализация модели и оценка качества аналитического решения;
* сохранение моделей;
* разработка и тестирование приложения для поддержки принятия решений специалистом предметной области, которое использовало бы найденную модель;

Данная работа позволила поработать не с учебными наборами данных, которые дают хорошо изученные решения, а с реальной производственной задачей.

Были выявлены возможные причины неудачи создания успешной модели:

* нечеткая постановка задачи, отсутствие дополнительной информации о зависимости признаков с точки зрения физики процесса (незначимые признаки являются для модели шумом, и мешают найти зависимость целевых от значимых входных признаков);
* исследование предварительно обработанных данных.

Возможно, на "сырых", не предобработанных данных можно было бы получить более качественные модели, воспользовавшись другими методами очистки и подготовки.

В ходе анализа итогов работы, были выявлены возможные варианты модернизации получившегося результата:

* углубление в структуру построения нейросетей;
* провести отбор признаков разными методами;
* после уменьшения размерности градиентный бустинг может улучшить свои результаты.

Даже на текущем этапе созданное веб-приложение показывает достаточную точность при выполнении результатов. А выбор в пользу разработки именно веб-приложения был сделан из-за их приемуществ.

Во-первых, веб-приложения предоставляют простоту использования.

Во-вторых, веб-приложения предоставляют большую гибкость. Они могут быть легко масштабируемы и модифицируемы.

В-третьих, веб-приложения предоставляют большую производительность. Они могут быть настроены таким образом, чтобы обрабатывать большие объемы данных и предоставлять быстрые ответы пользователям.

Наконец, веб-приложения предоставляют безопасность. Они могут быть настроены с различными уровнями безопасности, чтобы защитить данные и информацию от несанкционированного доступа.

Все эти преимущества делают веб-приложения привлекательными для многих пользователей, которые ищут простой, гибкий, производительный и безопасный способ доступа к информации и функциям.

# **Библиографический список**

1. Композиционные материалы : учебное пособие для вузов / Д. А. Иванов,
2. А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Издательство Юрайт, 2019 — 253 с. — (Высшее образование). — Текст : непосредственный.
3. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big
4. Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.
5. ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
6. Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html.
7. Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user.
8. Документация по библиотеке sklearn: – Режим доступа: https://scikitlearn.org/stable/user\_guide.html.
9. Руководство по быстрому старту в flask: – Режим доступа: https://flaskrussiandocs.readthedocs.io/ru/latest/quickstart.html.
10. Loginom Вики. Алгоритмы: – Режим доступа: https://wiki.loginom.ru/algorithms.html.
11. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/.
12. Yury Kashnitsky. Открытый курс машинного обучения. Тема 3.
13. Классификация, деревья решений: – Режим доступа: https://habr.com/ru/company/ods/blog/322534/.
14. Yury Kashnitsky. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес: – Режим доступа: https://habr.com/ru/company/ods/blog/324402/.
15. Alex Maszański. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest): – Режим доступа: https://proglib.io/p/mashinnoeobuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest
16. Alex Maszański. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга: – Режим доступа: <https://proglib.io/p/reshaemzadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga>.

**Приложение**

Github

Изображение выглядит как стол

Автоматически созданное описание