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 Pro»

Слушатель Панкратов Алексей Владимирович

Москва, 2024

Содержание

Введение

Выпускная квалификационная работа (ВКР) выполнена в рамках курса «Data Science Pro».

Тема работы: «Прогнозирование конечных свойств новых материалов (композиционных материалов).

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

Композиционные материалы находят широкое применение во многих отраслях промышленности. В связи с этим тема данной работы представляется весьма актуальной. Производство композитов связано с немалыми затратами. Даже если известны характеристики компонентов, невозможно точно рассчитать свойства готового композита. Поэтому для достижения желаемых характеристик необходимо провести множество испытаний различных сочетаний.

Чтобы сократить время и расходы на создание новых материалов, можно использовать систему поддержки производственных решений, основанную на принципах машинного обучения. Эта система поможет сократить количество испытаний и позволит пополнять базу данных новыми характеристиками материалов

1 Аналитическая часть

1.1 Постановка задачи

В ВКР исследуется композит с матрицей из базальтопластика и нашивками из углепластика. Исходный датасет содержит данные о свойствах матрицы и наполнителя, производственных параметрах и свойствах готового композита. Требуется разработать модели, прогнозирующие значения некоторых свойств в зависимости от значений остальных. Так же требуется разработать приложение, делающее удобным использование данных моделей специалистами предметной области.

Датасет состоит из двух файлов формата Excel: X_bp (составляющая из базальопластика) и X_nup (составляющая из углепластика).

Файл X_bp содержит:

– признаков: 10 и индекс;

– строк: 1023.

Файл X_nup содержит:

– признаков: 3 и индекс;

– строк: 1040.

В соответствии с заданием на ВКР необходимо осуществить объединение двух датасетов по индексу, тип объединения – INNER. Таким образом объединяются только те строки, для которых значения ключевых столбцов присутствуют в обоих исходных таблицах. После данного объединения часть строк из файла X_nup была отброшена. Дальнейшие исследования проводим с объединенным датасетом, содержащим 13 признаков и 1023 строк.

Описание признаков объединенного датасета приведено в таблице 1. Все признаки имеют тип float64, то есть вещественный. Пропусков в данных нет. Столбцов, полностью состоящих из уникальных значений в датасете не имеется. Все признаки, кроме «Угол нашивки», являются непрерывными,

количественными. «Угол нашивки» принимает только два целочисленных значения – 0 и 90.

Таблица 1 — Описание признаков датасета

Название признака	Файл	Тип данных	Непустых значений	Уникальных значений
Соотношение матрица-наполнитель	X_bp	float64	1023	1014
Плотность, кг/м3	X_bp	float64	1023	1013
Модуль упругости, ГПа	X_bp	float64	1023	1020
Количество отвердителя, м.%	X_bp	float64	1023	1005
Содержание эпоксидных групп,%_2	X_bp	float64	1023	1004
Температура вспышки, С_2	X_bp	float64	1023	1003
Поверхностная плотность, г/м2	X_bp	float64	1023	1004
Модуль упругости при растяжении, ГПа	X_bp	float64	1023	1004
Прочность при растяжении, МПа	X_bp	float64	1023	1004
Потребление смолы, г/м2	X_bp	float64	1023	1003
Угол нашивки, град	X_nup	int64	1023	2
Шаг нашивки	X_nup	float64	1023	989
Плотность нашивки	X_nup	float64	1023	988

1.2 Описание используемых методов

Предсказание значений вещественной, непрерывной переменной – это задача регрессии. Эта зависимая переменная должна иметь связь с одной или несколькими независимыми переменными, называемых также предикторами или

регрессорами. Регрессионный анализ помогает понять, как «типичное» значение зависимой переменной изменяется при изменении независимых переменных. В настоящее время разработано много методов регрессионного анализа. Например, простая и множественная линейная регрессия. Эти модели являются параметрическими в том смысле, что функция регрессии определяется конечным числом неизвестных параметров, которые оцениваются на основе данных.

1.2.1 Линейная регрессия

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1) и строится соответствующая прямая, известная как линия регрессии:

$$y = ax + b \tag{1}$$

Коэффициенты a и b, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов. Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид :

$$Y = b0 + b1 * x1 + b2 * x2 + \cdots + bn * xn, \tag{2}$$

где n - число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости. Линейная регрессия — первый тщательно изученный метод регрессионного анализа. Его главное достоинство — простота.

Такую модель можно построить и рассчитать даже без мощных вычислительных средств. Простота является и главным недостатком этого метода. Тем не менее, именно с линейной регрессии целесообразно начать подбор подходящей модели. На языке python линейная регрессия реализована в sklearn.linear_model.LinearRegression.

1.2.2 Лассо (LASSO)

Метод регрессии лассо (LASSO, Least Absolute Shrinkage and Selection Operator) — это вариация линейной регрессии, специально адаптированная для данных, которые имеют сильную сильную корреляцию признаков друг с другом. LASSO использует сжатие коэффициентов (shrinkage) и этим пытается уменьшить сложность данных, искривляя пространство, на котором они лежат. В этом процессе лассо автоматически помогает устранить или исказить сильно коррелированные и избыточные функции в методе с низкой дисперсией. Регрессия лассо использует регуляризацию L1, то есть взвешивает ошибки по их абсолютному значению.

Регуляризация позволяет интерпретировать модель. Если коэффициент стал 0, значит данный входной признак не является значимым. Этот метод реализованы в sklearn.linear_model.Lasso.

1.2.3 Метод опорных векторов

Метод опорных векторов (support vector machine, SVM) — один из наиболее популярных методов машинного обучения. Он создает гиперплоскость или набор гиперплоскостей в многомерном пространстве, которые могут быть использованы для решения задач классификации и регрессии. Чаще всего он применяется в постановке бинарной классификации. Основная идея заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Интуитивно, хорошее разделение достигается за счет гиперплоскости, которая имеет самое большое расстояние до ближайшей точки обучающей выборке

любого класса. Максимально близкие объекты разных классов определяют опорные вектора. Если в исходном пространстве объекты линейно неразделимы, то выполняется переход в пространство большей размерности. Решается задача оптимизации. Для вычислений используется ядерная функция, получающая на вход два вектора и возвращающая меру сходства между ними:

– линейная;

– полиномиальная;

– гауссовская (rbf).

Эффективность метода опорных векторов зависит от выбора ядра, параметров ядра и параметра C для регуляризации. Преимущество метода — его хорошая изученность. Недостатки:

– чувствительность к выбросам;

– отсутствие интерпретируемости.

Вариация метода для регрессии называется SVR (Support Vector Regression).

1.2.4 Случайный лес

Случайный лес (RandomForest) — представитель ансамблевых методов. Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив. Формула итогового решателя (3) — это усреднение предсказаний отдельных деревьев.

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x) \tag{3}$$

где N – количество деревьев; i – счетчик для деревьев; b – решающее дерево; x – сгенерированная нами на основе данных выборка.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

Преимущества случайного леса высокая точность предсказания, редко переобучается, практически не чувствителен к выбросам в данных, одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков, высокая параллелизуемость и масштабируемость. Из недостатков можно отметить, что его построение занимает больше времени. Так же теряется интерпретируемость. Метод реализован в sklearn.ensemble.RandomForestRegressor.

1.2.5 Нейронная сеть

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Структура нейронной сети пришла в мир программирования из биологии. Вычислительная единица нейронной сети — нейрон или персептрон. У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа. Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения. Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется:

– входной слой – его размер соответствует входным параметрам;

– скрытые слои – их количество и размерность определяет специалист;

– выходной слой – его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением. Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки

градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась. Для обновления весов в модели используются различные оптимизаторы. Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения. Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении.

1.2.6 Метрики качества моделей

Существует множество различных метрик качества, применимых для регрессии. В этой работе используются:

– R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то прогнозы сопоставимы по качеству с константным предсказанием;

– MSE (Mean Squared Error) или средне квадратичная ошибка. Применяется в случаях, когда требуется подчеркнуть большие ошибки и выбрать модель, которая дает меньше именно больших ошибок. Большие значения ошибок становятся заметнее за счет квадратичной зависимости;

– RMSE (Root Mean Squared Error) или корень из средней квадратичной ошибки принимает значениях в тех же единицах, что и целевая переменная. Метрика использует возведение в квадрат, поэтому хорошо обнаруживает грубые ошибки, но сильно чувствительна к выбросам;

– MAE (Mean Absolute Error) - средняя абсолютная ошибка так же принимает значениях в тех же единицах, что и целевая переменная;

– max error или максимальная ошибка данной модели в единицах измерения целевой переменной.

RMSE, MSE, MAE и max error принимают положительные значения. R2 в норме принимает положительные значения. Эту метрику надо максимизировать. Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

1.3 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.

Цели разведочного анализа данных:

1) понимание структуры и характеристик набора данных;

2) выявление аномалий и выбросов;

3) идентификация связей и корреляций между переменными;

4) подготовка данных для дальнейших этапов анализа.

Инструменты и методы разведочного анализа данных:

1) визуализация данных.

Визуализация данных позволяет нам увидеть и понять паттерны, тренды и взаимосвязи в данных через графику и диаграммы.

– гистограммы и диаграммы рассеяния;

– ящик с усами – это визуализация статистических характеристик распределения данных, таких как медиана, квартили и выбросы. Он помогает нам быстро оценить разброс и симметрию данных, а также выявить потенциальные аномалии;

– тепловая карта – это графическое представление матрицы данных, где цветовая шкала показывает степень взаимосвязи между переменными. Это помогает выявить паттерны и зависимости в больших наборах данных.

2) Сводные статистики и меры центральной тенденции. Это ключевые числовые метрики, которые помогают нам понять типичные и наиболее значимые значения в наборе данных:

– среднее (Mean): Это сумма всех значений, разделенная на количество значений. Оно представляет общую "среднюю" величину данных;

– медиана (Median): Это среднее значение двух средних значений, если количество значений четное, или среднее значение самого центрального числа, если количество значений нечетное;

– мода (Mode): Это значение, которое встречается наиболее часто в наборе данных. Мода может быть полезна для определения наиболее типичного значения.

3) Корреляционный анализ. Корреляционный анализ помогает нам понять, какие переменные взаимосвязаны между собой и насколько сильна эта связь. Коэффициент корреляции измеряет степень линейной зависимости между двумя переменными;

4) Преобразование данных (например, нормализация или стандартизация).

2 Практическая часть

2.1 Предобработка данных

Гистограммы распределения переменных и диаграммы «ящик с усами» приведены на рисунках 1-3.

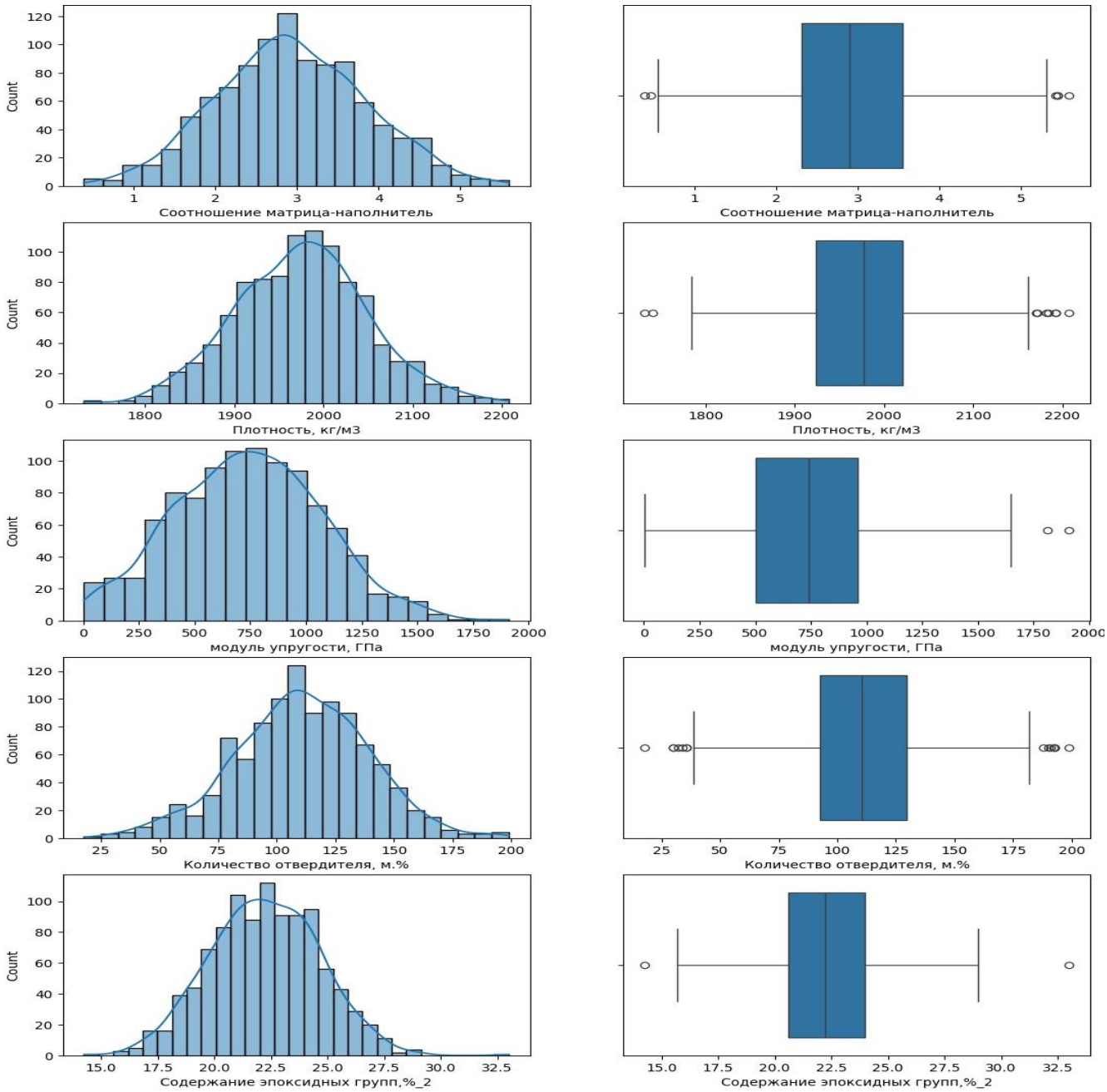

Рисунок 1– Гистограммы и диаграммы размаха признаков

Из рисунков видно что, все признаки, кроме «Угол нашивки», имеют

нормальное распределение и принимают неотрицательные значения.

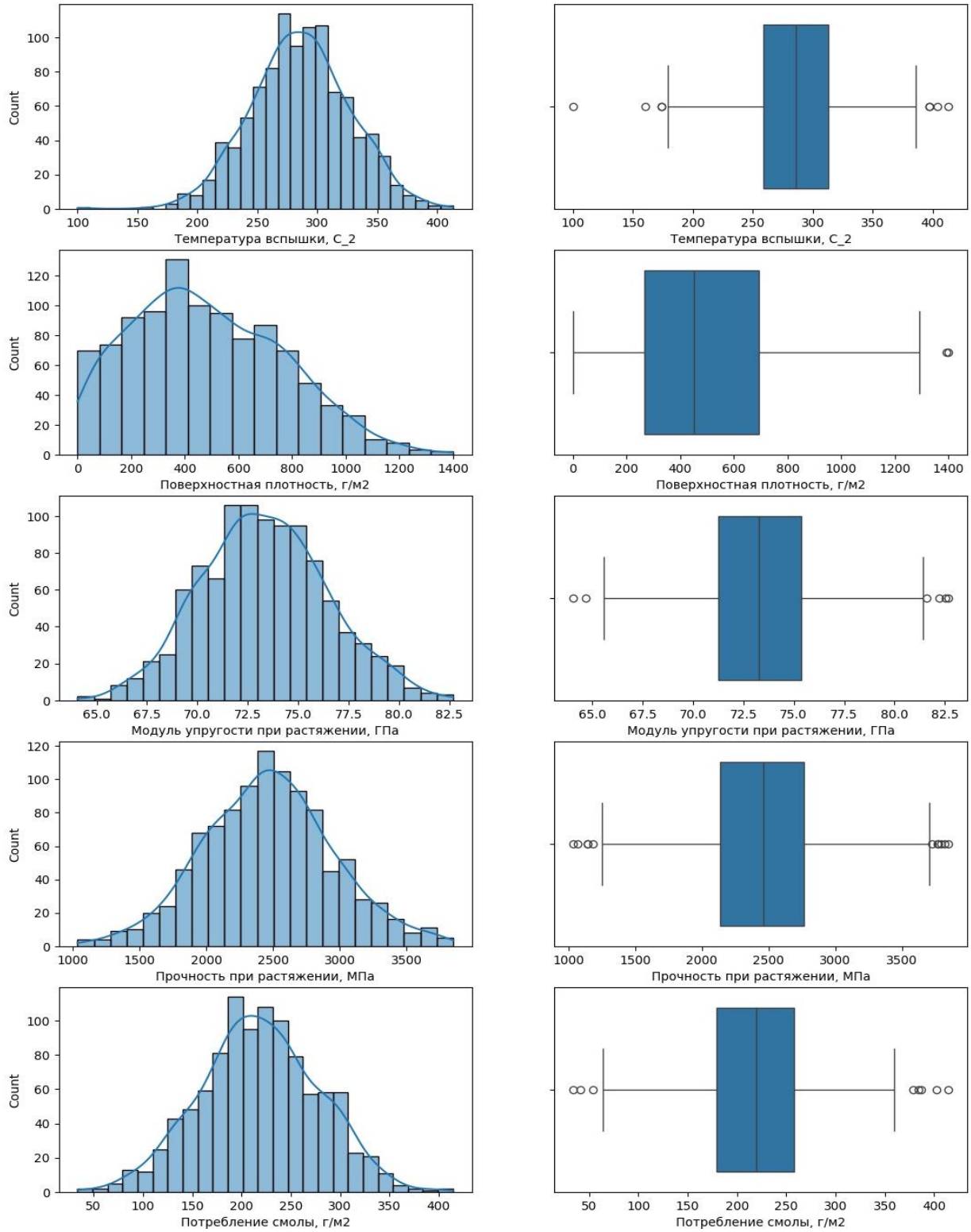

Рисунок 2 – Гистограммы и диаграммы размаха признаков

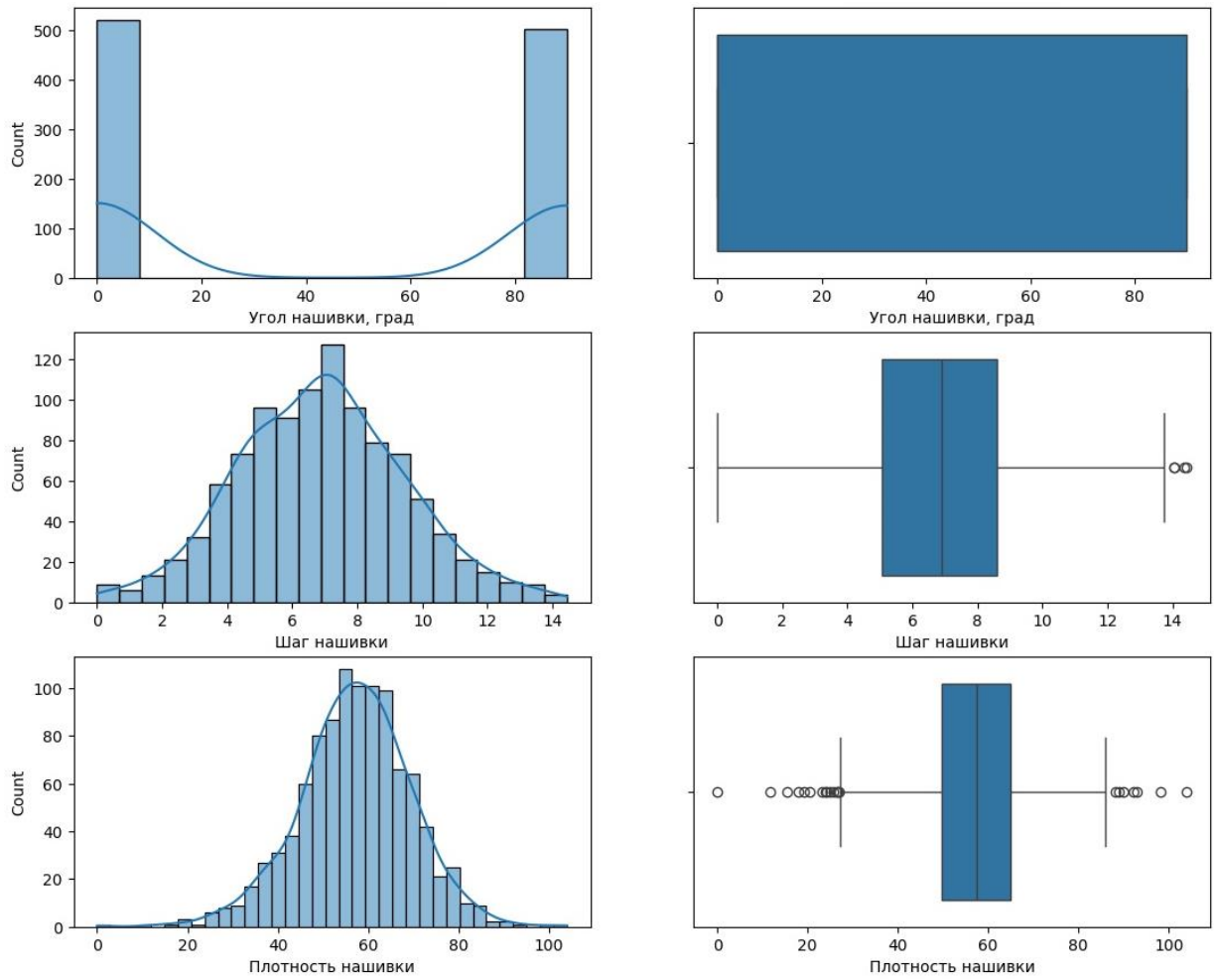

Рисунок 3 – Гистограммы и диаграммы размаха признаков

Отсутствие пропусков показывает, что датасет был возможно был предобработан. В “сырых” данных пропуски и значения некорректных типов, как правило, присутствуют.

Описательный анализ датасета, представленный в таблице 2, показал широкий разброс средних значений признаков, что говорит о необходимости нормализации данных.

Попарные графики рассеяния точек приведены на рисунке 4. По графикам рассеяния видно, что некоторые точки отстоят далеко от общего облака. Так визуально выглядят выбросы — аномальные, некорректные значения данных, выходящие за пределы допустимых значений признака.

Таблица 2 — Описательная статистика датасета

Признак	Среднее	Дисперсия	Минимум	Максимум
Соотношение матрица-наполнитель	2.930366	0.913222	0.389403	5.591742
Плотность, кг/м3	1975.734888	73.729231	1731.764635	2207.773481
Модуль упругости, ГПа	739.923233	330.231581	2.436909	1911.536477
Количество отвердителя, м.%	110.570769	28.295911	17.740275	198.953207
Содержание эпоксидных групп,%_2	22.244390	2.406301	14.254986	33.000000
Температура вспышки, С_2	285.882151	40.943260	100.000000	413.273418
Поверхностная плотность, г/м2	482.731833	281.314690	0.603740	1399.542362
Модуль упругости при растяжении, ГПа	73.328571	3.118983	64.054061	82.682051
Прочность при растяжении, МПа	2466.922843	485.628006	1036.856605	3848.436732
Потребление смолы, г/м2	218.423144	59.735931	33.803026	414.590628
Угол нашивки, град	44.252199	45.015793	0.000000	90.000000
Шаг нашивки	6.899222	2.563467	0.000000	14.440522
Плотность нашивки	57.153929	12.350969	0.000000	103.988901

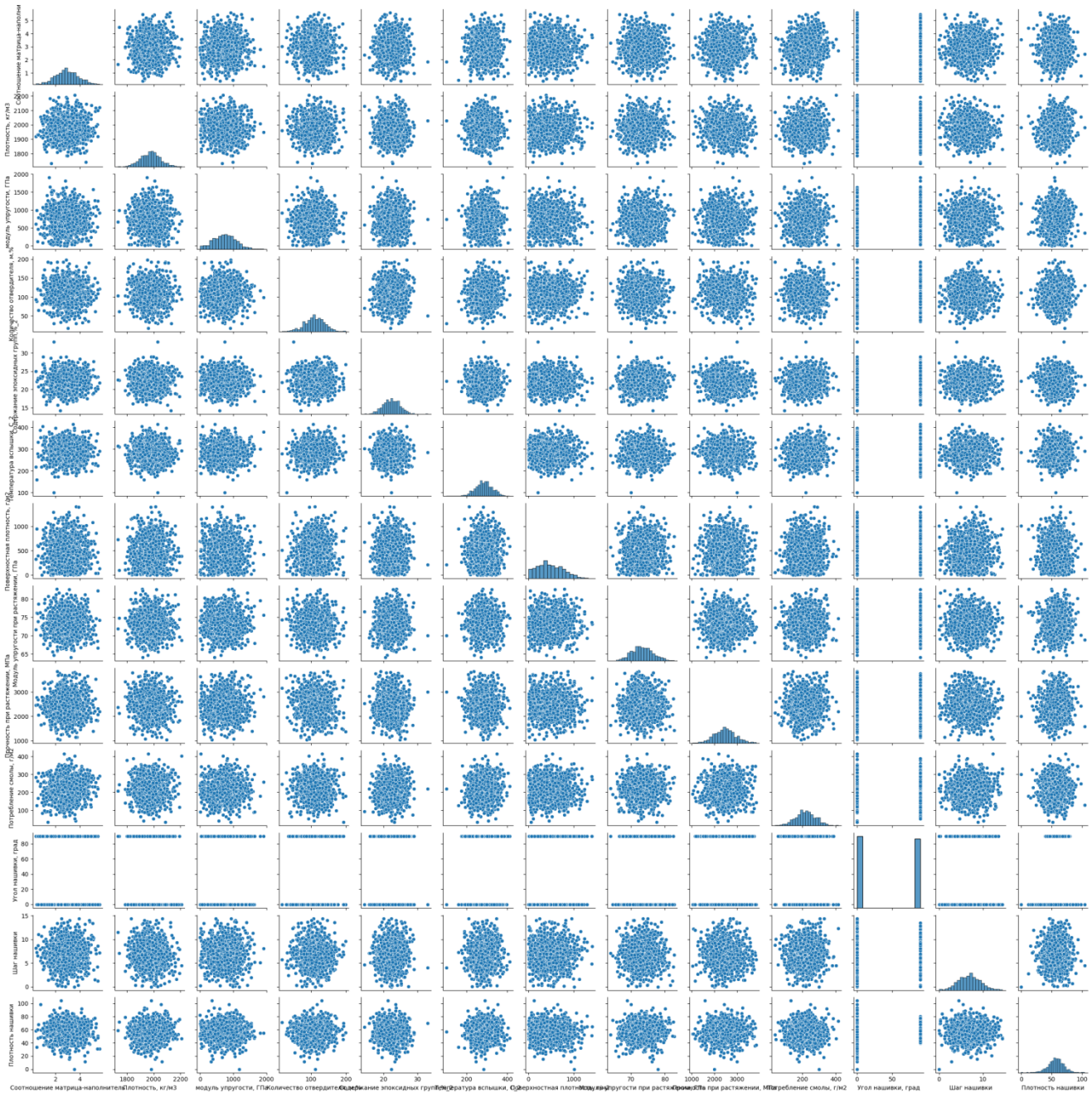

Рисунок 4 — Попарные графики рассеяния точек

На рисунке 4 представлен график попарного рассеяния точек не имеет явных зависимостей, которые станут основой работы моделей. Рассмотрим взаимосвязи между признаками с помощью матрицы-корреляции, приведенной на рисунке 5.

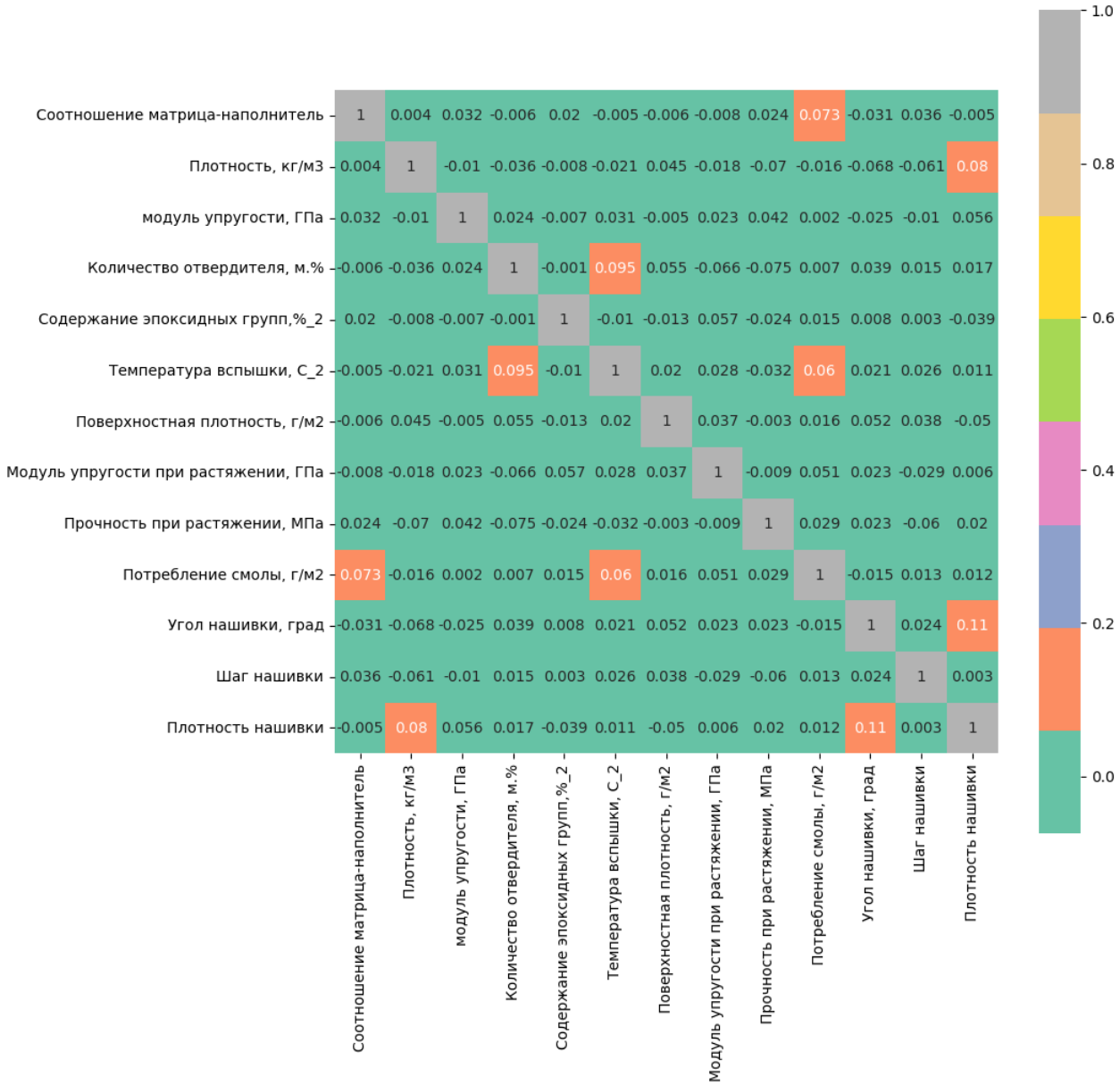

Рисунок 5 – Матрица корреляции

Из матрицы корреляции видно, что все коэффициенты корреляции близки к нулю, что означает отсутствие линейной зависимости между признаками.

Представленные на рисунках 4 и 5 графики попарного рассеяния точек и матрицы-корреляции показываю отсутствие линейных зависимостей между исследуемыми параметрами. Максимальное значение коэффициента корреляции для

«Модуля упругости при растяжении» с признаком «Количество отвердителя» -0,07, для «Прочности при растяжении» с признаком «Количество отвердителя» -0,08.

Существуют методы выявления выбросов для признаков с нормальным распределением: метод 3-х сигм; метод межквартильных расстояний.

Поскольку известно, что датасет очищен от явного шума, то воспользуемся тремя наборами данных. Первый датасет (X1) оставим исходный без изменений, для получения второго (X2) применим метод межквартильных расстояний. Значения, определенные как выбросы, удаляем. После этого осталось в втором датасете осталось 936 строк и 13 признаков-переменных. В задании целевыми переменными указаны: – модуль упругости при растяжении, Гпа; – прочность при растяжении, МПа; – соотношение матрица-наполнитель. В третьем датасете (X3) оставим переменные, корреляция которых с целевыми не ниже 0.03. К ним относятся:

– количество отвердителя;

– содержание эпоксидных групп;

– поверхностная плотность;

– модуль упругости при растяжении;

– прочность при растяжении;

– потребление смолы;

– шаг нашивки.

2.2 Разработка и обучение моделей

В качестве исследуемых выбраны следующие модели:

– линейная регрессия;

– модель Lasso;

– модель SVR;

– модель случайного леса.

Стратегия исследования предполагала построение каждой модели с параметрами по умолчания для каждого набора данных. Выбор наилучшего

набора данных с точки зрения значений метрик качества. Далее, для выбранного набора данных построение моделей с поиском оптимальных параметров. Выбор наилучшей из них. Таким образом, в таблицах 3-5 представлены результаты моделирования прогноза модуля упругости при растяжении:

Таблица 3 – Результаты моделирования на датасете X1

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.034	9.671	3.101	2.486	7.971
Lasso	-0.020	9.533	3.080	2.474	7.975
SVR	-0.059	9.896	3.137	2.541	7.892
RandomForestRegressor	-0.088	10.153	3.179	2.534	8.301

Таблица 4 – Результаты моделирования на датасете X2

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.043	9.403	3.064	2.445	7.578
Lasso	-0.016	9.177	3.026	2.420	7.260
SVR	-0.039	9.382	3.058	2.448	7.429
RandomForestRegressor	-0.061	9.587	3.090	2.469	7.781

Таблица 5 – Результаты моделирования на датасете X3

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.015	9.321	3.038	2.447	7.480
Lasso	-0.016	9.394	3.046	2.451	7.208
SVR	-0.071	9.779	3.114	2.491	7.568
RandomForestRegressor	-0.094	9.988	3.147	2.524	7.846

Анализ результатов показал, что ни одна из выбранных мной моделей не оказалась подходящей для представленных данных. Коэффициент детерминации R2 близок к 0 для линейных моделей и метода Lasso и опорных векторов. Наилучшие показатели качества получены при обучение моделей на датасете X2. Поэтому, на этом датасете был произведен выбор модели регрессии и подбор параметров с помощью функции GridSearchCV(model, params, cv, scoring). В результате моделирования получены следующие значения показателей качества регрессионных моделей (таблица 6).

Таблица 6 – Результаты моделирования на датасете X2 с подбором параметров.

Вид модели	R2	MSE	RMSE	MAE	max_ error
Lasso(alpha=0.1)	-0.016	9.177	3.026	2.420	7.260
SVR(C=5, kernel='sigmoid')	-0.014	9.152	3.021	2.413	7.474
RandomForestRegressor(criterion='absolute_error', max_depth=4, max_features=1, random_state=42)	-0.013	9.148	3.021	2.416	7.368

Исходя из полученных данных, наилучшей регрессионной моделью выступает модель на основе опорных векторов (SVR) с параметрами C=5, kernel='sigmoid'.

Аналогичные исследования были проведены для выбора и обучения алгоритма машинного обучения для параметра «Прочность при растяжении». Полученные результаты представлены в таблицах 7-9.

Таблица 7 – Результаты моделирования на датасете X1

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.028	224429.643	470.898	371.614	1205.611
Lasso	-0.018	222509.090	468.784	369.669	1195.738
SVR	-0.004	219749.473	465.802	366.656	1179.055
RandomForestRegressor	-0.070	232331.557	479.606	381.433	1190.454

Таблица 8 – Результаты моделирования на датасете X2

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.058	223584.571	471.642	376.792	1121.774
Lasso	-0.050	222075.214	470.022	374.775	1121.612
SVR	-0.027	218386.233	465.780	370.110	1124.890
RandomForestRegressor	-0.067	224920.171	473.219	377.493	1115.646

Таблица 9 – Результаты моделирования на датасете X3

Вид модели	R2	MSE	RMSE	MAE	max_error
LinearRegression	-0.022	204439.568	451.356	360.134	1156.106
Lasso	-0.020	203966.347	450.816	359.889	1155.016
SVR	-0.019	203832.817	450.657	359.918	1143.236
RandomForestRegressor	-0.061	212276.896	459.622	370.849	1201.386

Анализ результатов показал, что наилучшие показатели качества получены при обучение моделей на датасете X3. Поэтому, на этом датасете был произведен выбор модели регрессии и подбор параметров с помощью функции GridSearchCV(model, params, cv=cv, scoring=scoring). В результате моделирования получены следующие значения показателей качества регрессионных моделей (таблица 10).

Таблица 10 – Результаты моделирования на датасете X3 с подбором параметров.

Вид модели	R2	MSE	RMSE	MAE	max_ error
Lasso(alpha=1)	-0.020	203966.347	450.816	359.889	1155.016
SVR(C=0.01, kernel='poly')	-0.018	203692.791	450.509	359.800	1143.599
RandomForestRegressor(bootstrap=False, criterion='absolute_error',max_depth=4, max_features=1, n_estimators=50, random_state=42)	0.002	199806.448	446.061	356.336	1142.427

Исходя из полученных данных, наилучшей регрессионной моделью выступает модель “случайного леса” (RandomForestRegressor) с параметрами bootstrap=False, criterion='absolute_error', max_depth=4, max_features=1, n_estimators=50, random_state=42

2.3 Тестирование моделей

Обучение и тестирование производилось для двух признаков – модуля упругости и прочности при растяжении.

Модель для прогноза значения модуля упругости – SVR(C=5, kernel='sigmoid'). Значения метрик качества модели на обучающем и тестовом наборе данных представлены в таблице 11.

Таблица 11 – Метрики качества модели SVR (C=5, kernel='sigmoid') на датасете X2

Набор данных	R2	MSE	RMSE	MAE	max_error
обучающий	-0.001	9.362	3.060	2.522	7.873
тестовый	-0.042	9.746	3.122	2.558	9.327

Из таблицы видно, что модель на тестовом наборе показала чуть хуже результаты, чем на обучающей выборке. Но в целом, значения не сильно

различаются. Коэффициент детерминации R2 имеет отрицательные значения, что указывает на нецелесообразности использования данной модели.

На рисунке 6 представлены результаты моделирования.

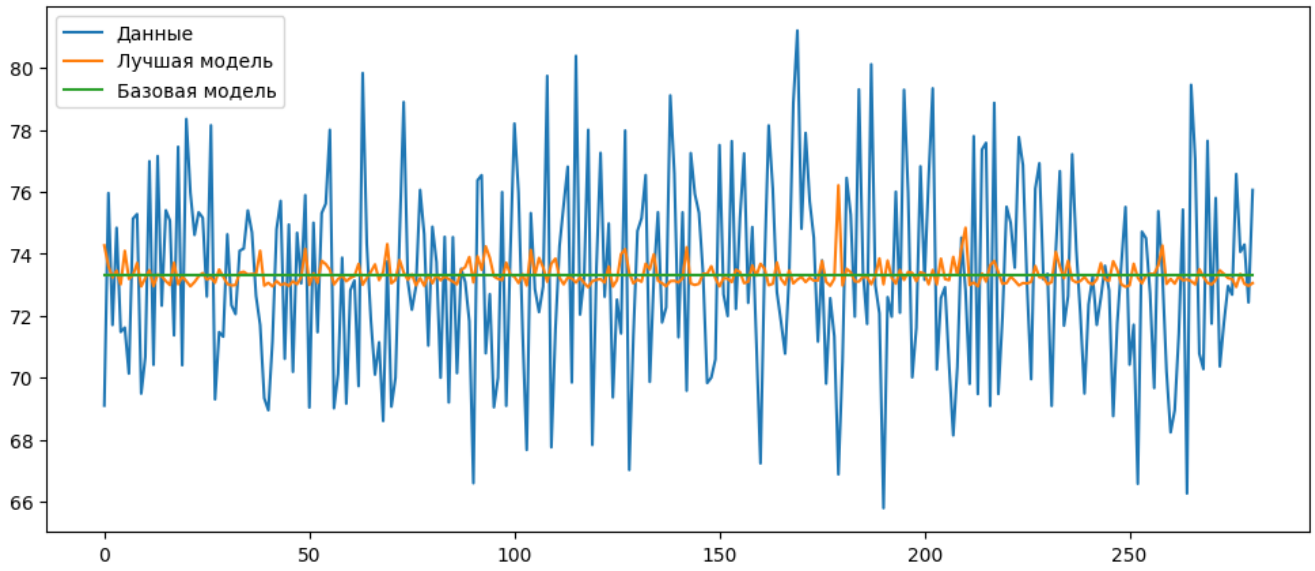

Рисунок 6 – Визуализация работы модели SVR()

В качестве базовой модели взят DummyRegressor, возвращающий среднее значение целевого признака.

Модель для прогноза значения прочности при растяжении – RandomForestRegressor(bootstrap=False, criterion='absolute_error',max_depth=4, max_features=1, n_estimators=50, random_state=42). Значения метрик качества модели на обучающем и тестовом наборе данных представлены в таблице 12.

Таблица 12 – Метрики качества модели RandomForestRegressor на датасете X3

Набор данных	R2	MSE	RMSE	MAE	max_error
обучающий	0.006	237638.567	487.482	389.470	1159.809
тестовый	-0.020	240801.955	490.716	390.868	1173.772

Аналогично предыдущей, модель на тестовом наборе показала чуть хуже результаты, чем на обучающей выборке. Значения не сильно различаются. Коэффициент детерминации R2 около 0, что свидетельствует о низкой прогнозной способности модели.

На рисунке 7 представлены результаты моделирования.

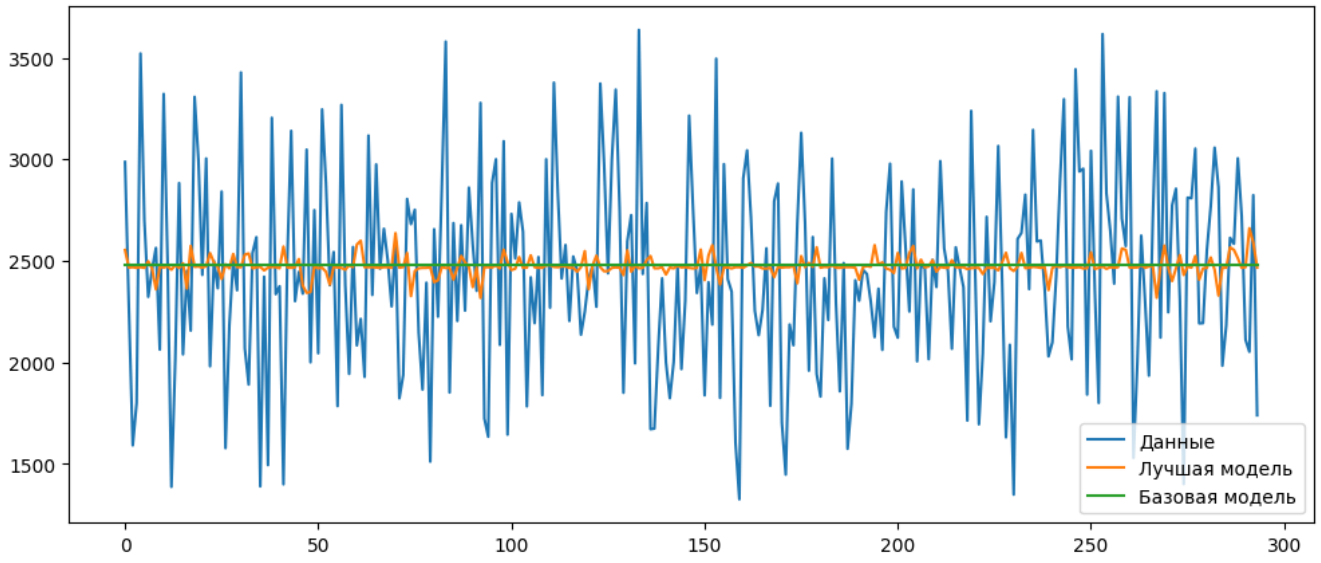

Рисунок 7 – Визуализация работы модели RandomForestRegressor (…)

В качестве базовой модели взят DummyRegressor, возвращающий среднее значение целевого признака.

2.4 Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель

Для решения этой задачи воспользуемся методами библиотек sklear и tensorlow. Для сравнения также понадобится базовая модель DummyRegressor, возвращающая среднее целевого признака.

2.4.1 MLPRegressor из библиотеки sklearn

Ключевые параметры MLPRegressor включают hidden_layer_sizes (количество нейронов в каждом скрытом слое), activation (функция активации) и solver (алгоритм оптимизации).

Путем варьирования значениями параметров hidden_layer_sizes [24, 96], activation [‘logistic’, ‘relu’] и solver [‘sgd’, ‘adam’] получена нейронная сеть со следующей архитектурой:

– количество скрытых слоев: 8;

– нейронов на каждом слое: 72;

– функция активации: relu;

– оптимизатор: adam, (решатель для оптимизации веса);

– пропорция разбиения данных на тестовые и валидационные: 30%;

– количество итераций: 5000.

Таким образом, сформирована модель MLPRegressor(early_stopping=True, hidden_layer_sizes=(72, 72, 72, 72, 72, 72, 72, 72), max_iter=5000, random_state=42, validation_fraction=0.3, verbose=True).

Обучение модели происходило на датасете X4 за 30 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE). На рисунке 8 представлена кривая потерь при обучении модели.

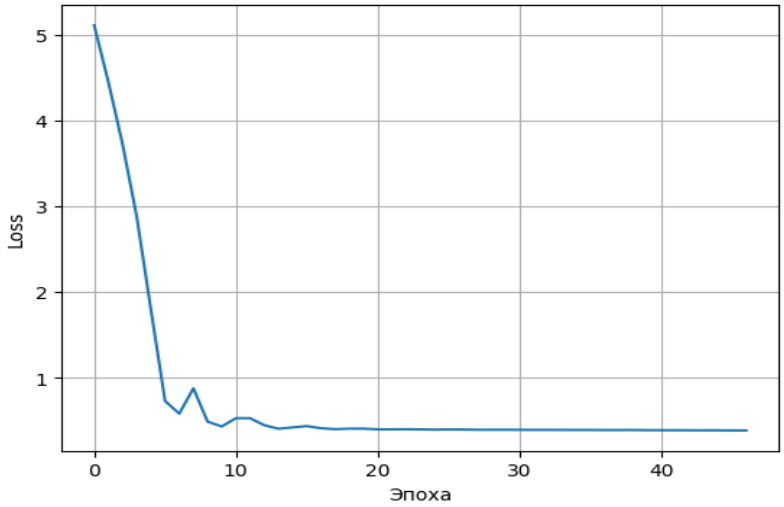

Рисунок 8 - Функция потерь

Значения метрик качества модели на обучающем и тестовом наборе данных представлены в таблице 13.

Таблица 13 – Метрики качества моделей DummyRegressor и MLPRegressor

Набор данных	R2	MSE	RMSE	MAE	max_error
DummyRegressor	-0.000	0.861	0.928	0.743	2.539
MLPRegressor	0.003	0.858	0.926	0.743	2.506

Сравнительный анализ показывает, что качество модели на уровне простой базовой регрессионной модели. Визуализация результатов, полученных нейросетью, приведены на рисунке 9.

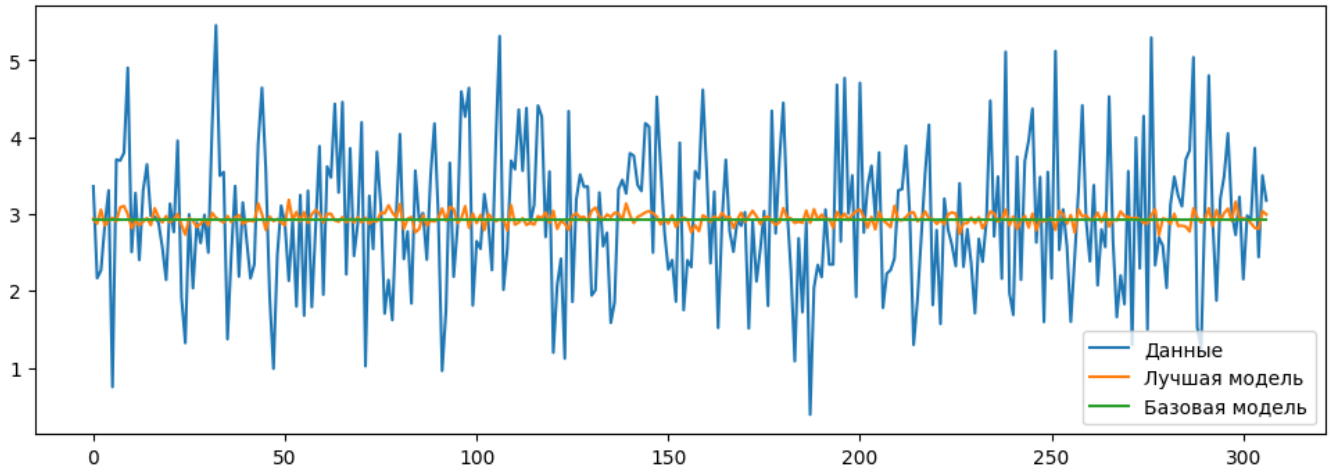

Рисунок 9 - Визуализация результатов MLPRegressor

2.4.2 Нейросеть из библиотеки tensorflow

Аналогично предыдущей нейросети, путем варьирования значениями параметров keras.layers.Dense [1,8], units [12, 96] и activation [‘logistic’, ‘relu’] получена нейронная сеть со следующей архитектурой:

– входной слой для 12 признаков;

– выходной слой для 1 признака;

– скрытых слоев: 8;

– нейронов на каждом скрытом слое: 7 слоев по 12, 8 слой 4 нейрона;

– активационная функция скрытых слоев: relu;

– оптимизатор: adam;

– loss-функция: MeanSquaredError.

Обучение модели происходило на датасете X1_ за 30 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE). На рисунке 10 представлена кривая MSE при обучении модели.

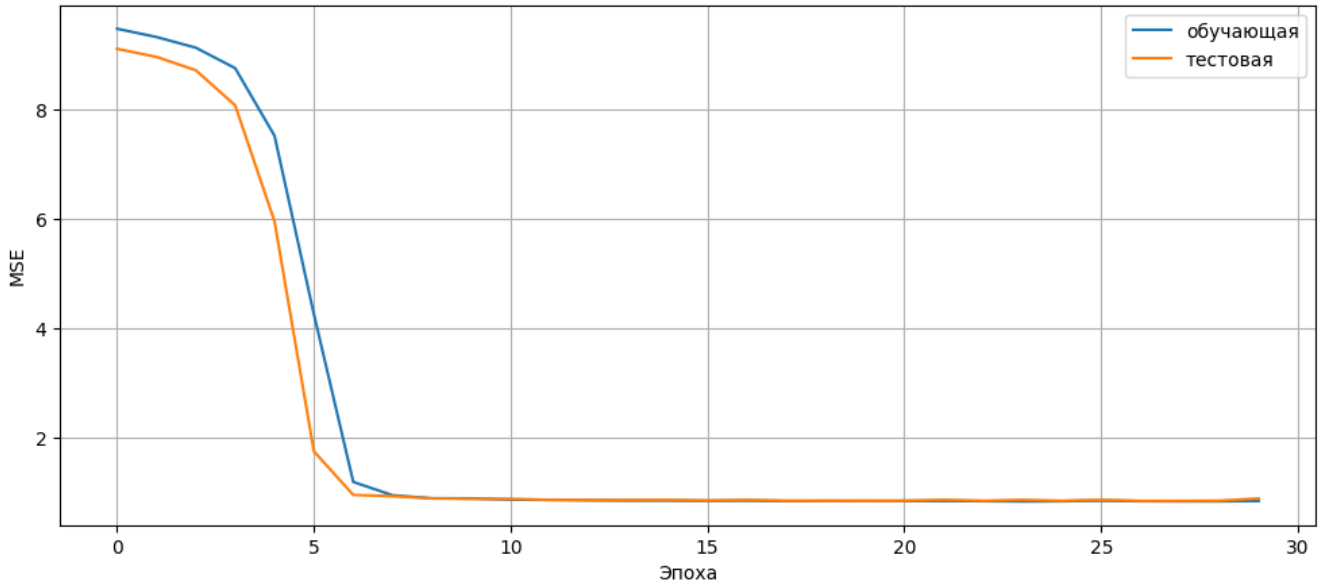

Рисунок 10 – Зависимость MSE от эпохи

Значения метрик качества модели на обучающем и тестовом наборе данных представлены в таблице 14.

Таблица 14 – Метрики качества моделей DummyRegressor и keras

Набор данных	R2	MSE	RMSE	MAE	max_error
DummyRegressor	-0.013	0.831	0.911	0.743	2.419
keras	-0.048	0.902	0.950	0.765	2.773

Сравнительный анализ показывает, что качество модели на уровне простой базовой регрессионной модели.

2.5 Разработка приложения.

В интерфейсе использованы наименования параметров из датасета.

После запуска приложения для получения прогноза модуля упругости при растяжении и прочности при растяжении - введите "1". Для того чтобы получить рекомендацию соотношения матрица-наполнитель - введите "2". Для завершения работы приложения - введите "3" (или "выход", или "exit", или пустую строку).

Пример ввода параметров приложения для получения прогноза модуля упругости при растяжении и прочности при растяжении представлен на рисунке 11. В качестве параметров внесены значения близкие к среднему и в рамках дисперсии. Результаты расчета представлены на рисунке 12.

```
    Для получения прогноза модуля упругости при растяжении и
    прочности при растяжении введите исходные 11 параметров композитного материала:

Введите Соотношение матрица-наполнитель:  2.9
Введите Плотность, кг/м3:  1975
Введите модуль упругости, ГПа:  739
Введите Количество отвердителя, м.%:  110
Введите Содержание эпоксидных групп,%_2:  22
Введите Температура вспышки, С_2:  285.8
Введите Поверхностная плотность, г/м2:  482
Введите Потребление смолы, г/м2:  218.4
Введите Угол нашивки, град:  44.25
Введите Шаг нашивки:  6.9
Введите Плотность нашивки:  57

Проводится прогноз целевых параметров ['Модуль упругости при растяжении, ГПа', 'Прочность при растяжении, МПа']
                                     1
Соотношение матрица-наполнитель     2.90
Плотность, кг/м3                 1975.00
модуль упругости, ГПа             739.00
Количество отвердителя, м.%       110.00
Содержание эпоксидных групп,%_2    22.00
Температура вспышки, С_2          285.80
Поверхностная плотность, г/м2     482.00
Потребление смолы, г/м2           218.40
Угол нашивки, град                 44.25
Шаг нашивки                         6.90
Плотность нашивки                  57.00
```

Рисунок 11 – Пример ввода параметров приложения для расчета прогноза «Модуля упругости при растяжении» и «Прочности при растяжении»

```
Прогнозное значение показателя Модуль упругости при растяжении, ГПа = 720.8411 ГПа
Прогнозное значение показателя Прочность при растяжении, МПа = 2404.5437 МПа
```

Рисунок 12 – Результаты расчета прогноза «Модуля упругости при растяжении» и «Прочности при растяжении»

2.6 Создание удаленного репозитария

На GitHub создана страница слушателя.

Созданный репозиторий: https://github.com/Servkrut/Data_science.

Заключение

В ходе выполнения ВКР были изучены способы анализа и предобработки данных. Построенные модели показали, что исходный датасет является предобработанным и не содержит реальных значений для отработки обучения и тренировки моделей.

Для большинства линейных моделей коэффициент детерминации R2 имеет отрицательные значения, что указывает на нецелесообразности использования данных моделей.

Разработанная модель нейронной имеет значение коэффициента детерминации близким к нулю и принимает значения близкие к базовой.

К сожалению, в результате исследований не удалось получить модели, которые бы описывали закономерности предметной области. Дальнейшее направление работы в данном направлении вижу в попытке использовать методы уменьшения размерности, например метод главных компонент, углубиться в изучение нейросетей, попробовать различные архитектуры, параметры обучения, проконсультироваться у экспертов в предметной области. Также возможные не высокие результаты являются следствием недостатка знания и отсутствие опыта в данной области.

Библиографический список

1 Композиционные материалы : учебное пособие для вузов / Д. А. Иванов, А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Издательство Юрайт, 2019 — 253 с. — (Высшее образование). — Текст : непосредственный.

2 Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.

3 ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.

4 Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html.

5 Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user.

6 Документация по библиотеке pandas: – Режим доступа: https://pandas.pydata.org/docs/user_guide/index.html#user-guide.

7 Документация по библиотеке matplotlib: – Режим доступа: https://matplotlib.org/stable/users/index.html.

8 Документация по библиотеке seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial.html.

9 Документация по библиотеке sklearn: – Режим доступа: https://scikit-learn.org/stable/user_guide.html.

10 Документация по библиотеке keras: – Режим доступа: https://keras.io/api/.

11 Руководство по быстрому старту в flask: – Режим доступа: https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html.

12 Loginom Вики. Алгоритмы: – Режим доступа: https://wiki.loginom.ru/algorithms.html.

13 Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа:https://habr.com/ru/company/vk/blog/513842/.

14 Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19.

15 Yury Kashnitsky. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей: – Режим доступа: https://habr.com/ru/company/ods/blog/322534/.

16 Yury Kashnitsky. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес: – Режим доступа: https://habr.com/ru/company/ods/blog/324402/.

17 Alex Maszański. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest): – Режим доступа: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12.

18 Alex Maszański. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга: – Режим доступа: https://proglib.io/p/reshaem-zadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga-2021-11-25.