МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования
УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
имени первого Президента России Б.Н. Ельцина

ИНСТИТУТ ЕСТЕСТВЕННЫХ НАУК И МАТЕМАТИКИ

Кафедра магнетизма и магнитных наноматериалов

Валидация модели машинного обучения для прогнозирования магнитных свойств нанокристаллических сплавов типа FINEMET

Направление подготовки 27.04.01 «Стандартизация и метрология»

Образовательная программа «Метрологическое обеспечение научных исследований и наукоемких технологий»

Магистерская диссертация

Зав. кафедрой:
д.ф.-м.н., проф. В.О. Васьковский

Степановой
Ксении Александровной

Нормоконтролер:
к.ф.-м.н., Е.А. Степанова

Научные руководители:
к.ф.-м.н., В.А. Катаев
к.ф.-м.н., А.С. Болячкин

Екатеринбург
2022

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования

УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

имени первого Президента России Б.Н. Ельцина

ИНСТИТУТ ЕСТЕСТВЕННЫХ НАУК И МАТЕМАТИКИ

УТВЕРЖДАЮ

Руководитель образовательной программы

Васьковский В. О. (_____)

« 07 » февраля 2022 г.

Код, наименование направления: 27.04.01 «Стандартизация и метрология»

Наименование программы: Метрологическое обеспечение научных исследований и наукоёмких технологий

Группа: МЕНМ – 202703

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Степановой Ксении Александровны
(Фамилия, имя, отчество)

Квалификация: магистр
(бакалавр, специалист, магистр)

Провести научное исследование по теме: Валидация модели машинного обучения для прогнозирования магнитных свойств нанокристаллических сплавов FINEMET

Срок представления работы научному руководителю: «___» _____ 20___ г.

Научный руководитель _____ (Катаев В. А.)
(Подпись) (И.О. Фамилия)

Задание принял к исполнению _____
(Подпись)

РЕФЕРАТ

Тема магистерской диссертации: «Валидация модели машинного обучения для прогнозирования магнитных свойств нанокристалических сплавов типа FINEMET». Работа изложена на 61 страницах, включает 7 таблиц, 42 рисунка, содержит 40 использованных источников, 1 приложение.

КЛЮЧЕВЫЕ СЛОВА: МАШИННОЕ ОБУЧЕНИЕ, ВАЛИДАЦИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ, СЛУЧАЙНЫЙ ЛЕС, К – БЛИЖАЙШИЕ СОСЕДИ, МЕТОД ОПОРНЫХ ВЕКТОРОВ

В работе была произведена разработка модели машинного обучения на языке программирования Python, а также проведена ее валидация на этапах жизненного цикла.

Целью создания модели машинного обучения является прогнозирование магнитных свойств нанокристаллических сплавов на основе железа по химическому составу и условиям обработки.

Процесс валидации модели машинного обучения позволяет не только произвести контроль за соблюдением требований, предъявляемых при разработке и эксплуатации модели, к результатам, полученных с помощью моделирования, но и способствует внедрению модели в процесс производства.

Процесс валидации включал в себя:

Валидация данных, в ходе которой были оценены типы, пропуски данных, соответствие цели исследования, распределения признаков и целевых характеристик, изучены корреляции признаков и целевых характеристик.

Валидация алгоритмов, применяемых в модели: были проанализированы параметры алгоритмов с целью соблюдения требования о корректной обобщающей способности модели (отсутствие недо- и переобучения).

Оценка работы модели, благодаря которой был произведен анализ полученных результатов с помощью тестовых данных.

Верификация результатов с помощью актуальных данных, полученных из статей, опубликованных с 2010 по 2022 год.

В результате валидации модели было показано высокое качество разработанной модели, позволяющее получить оценки качества R^2 0,65 и выше.

The magister dissertation's theme is "Validation of machine learning model to predict magnetic properties of nanocrystalline FINEMET type alloys". The work consists of 61 pages, 7 tables, 42 pictures, 40 literature sources.

KEYWORDS: MACHINE LEARNING, MACHINE LEARNING MODEL VALIDATION, RANDOM FOREST, K – NEAREST NEIGHBORS, SUPPORT VECTOR REGRESSOR,

In this work machine learning model was developed by Python programming language, and also was validated at stages of model's life cycle.

The purpose of creating the machine learning model is to predict the magnetic properties of Fe-based nanocrystalline alloys by chemical composition and processing conditions.

The validation of machine learning models allows not only to control the requirements for development and operation of the models, for the results obtained by modeling, but also contributes to the introduction of the model into production process.

The validation process included:

Data validation: data types and omissions, compliance with the purpose of the study, distribution of features and target characteristics were evaluated, correlations of features and target characteristics were studied.

Algorithms validation: the parameters of the algorithms were analyzed in order to comply with the requirement for the correct generalizing ability of the model (without under- and overfitting).

Evaluation of the model work: the analysis of the obtained results was carried out using test data.

Verification of results using actual data obtained from articles published since 2010 to 2022.

As a result of the model validation, the high quality of the developed model was shown, which makes it possible to obtain quality metric R^2 0.65 and higher.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 8

1 Обзор литературы 10

1.1 Влияние химического состава и условий термической обработки на магнитные свойства 10

1.2 Машинное обучение 11

1.3 Жизненный цикл модели машинного обучения 12

1.4 Валидация 13

2 Постановка задачи 19

3 Методика 20

3.1 Сбор данных 20

3.2 Анализ данных 21

3.3 Отбор важных признаков 30

3.4 Выбор метрики оценки алгоритма 33

3.5 Выбор алгоритмов 35

3.6 Оптимизация параметров 37

4 Результаты и их обсуждение 47

ЗАКЛЮЧЕНИЕ 58

СПИСОК ЛИТЕРАТУРЫ 60

ПРИЛОЖЕНИЕ А 64

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящей работе применяют следующие термины и определения.

Алгоритм – это программный код, состоящий из набора последовательных инструкций, при выполнении которых достигается результат.

Валидация - подтверждение путем проверки и предоставления объективных доказательств выполнения особых требований к конкретному предусмотренному применению, а также того, что все требования выполняются надлежащим образом и в полном объеме, и что обеспечивается прослеживание выполнения системных требований.

Машинное обучение – класс методов искусственного интеллекта, изучающий методы построения алгоритмов для обучения моделей машинного обучения.

Модель машинного обучения – совокупность методов искусственного интеллекта для автоматизированного прогноза целевой характеристики на основе опытных данных.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящей работе применяют следующие обозначения и сокращения.

γ – перекос данных

μ – магнитная проницаемость

H_c – коэрцитивная сила

KNN (от англ. k - Nearest Neighbors) – метод ближайших соседей

MLP (от англ. Multi-Layer Perceptron) - многослойный персептрон. Нейронная сеть, состоящая из входного, выходного и расположенных между ними одного или нескольких скрытых слоев нейронов

n – количество образцов с указанной характеристикой

Random Forest – случайный лес

SVR (от англ. Support Vector Regressor) – метод опорных векторов

ВВЕДЕНИЕ

В XXI веке человечество стоит на пороге Четвертой промышленной революции, так называемой «Индустрия 4.0». Основа этой революции состоит в цифровизации во всех сферах человеческой деятельности. Благодаря новым технологиям «Индустрия 4.0» выводит автоматизацию, мониторинг и анализ на новый уровень. Неотъемлемой частью технологий является и инфраструктура качества.

Метрологическое обеспечение является гарантом качества технологий «Индустрии 4.0». Развитие метрологии в этих сферах является актуальной задачей в этой промышленной революции.

Поскольку «Индустрия 4.0» предполагает сбор большого объема данных из широкого спектра источников - большой интерес представляет метрология в анализе больших объемов данных «Big Data». Целью метрологии в этой сфере является метрологическая оценка методов анализа больших объемов данных и машинного обучения.

Машинное обучение уже сегодня применяется во множестве сфер деятельности человека: от торговли и производства до биотехнологий. Одной из таких сфер является материаловедение.

Модификация свойств материалов происходит методом проб и ошибок. Использование машинного обучения в области разработки новых материалов может значительно уменьшить расход ресурсов.

Аморфные и нанокристаллические сплавы на основе железа широко применяются в радио- и электропромышленности, так как свойства эти материалов превосходят классические магнитомягкие материалы. К таким сплавам относится нанокристаллические сплавы на основе железа FINEMET. Совершенствование свойств сплавов приведет к расширению их функционала и применения.

Раджеш и соавт. [1] представили комбинированный подход CALPHAD и машинного обучения для прогнозирования размера и объемной доли нанокристаллов на основе входных данных о составе и условиях термообработки.

В работе [2] Уэнг и соавт. сосредоточились на предсказании магнитных свойств на основе экспериментальных данных, извлеченных из литературы.

В работе [3] авторы предсказывали индукцию магнитного насыщения при помощи машинного обучения, учитывая показатели формирования аморфной фазы и других термодинамических, кинетических и структурных свойств.

Также существуют работы по совершенствованию и классических аморфных составов на основе железа путем машинного обучения [4, 5]. Целью этих исследований являлось

предсказание термостабильности и индукции насыщения, а также улучшение механических свойств: прочности и пластичности.

Для непосредственного внедрения модели машинного обучения в процесс изготовления сплавов необходимо производить проверки работы модели на различных этапах жизненного цикла модели. Такой проверкой может служить валидация. Модель должна быть нацелена на точный прогноз с минимальной долей погрешности. По результатам валидации можно сделать выводы о правильности работы модели, ее пригодности для решения задач и корректности ее выводов.

Таким образом, валидация модели поможет в конечном счете достичь основной цели разработки – внедрение в процесс разработки новых составов.

1 Обзор литературы

1.1 Влияние химического состава и условий термической обработки на магнитные свойства

В состав классического сплава FINEMET входят химические элементы: Fe, Cu, Nb, Si, B. Каждый из них выполняет свою функцию:

- Кремний и бор используют для получения аморфной структуры;
- Медь способствует формированию кластеров, обогащенных медью на начальной стадии отжига, и обеспечивает начало кристаллизации из большого числа центров по всему объему материала;
- Ниобий формирует более мелкие кластеры меди и сдерживает рост кристаллической фазы.

В результате нанокристаллизации формируются зерна $Fe_{80}Si_{20}$ с ОЦК решеткой и размером ~10 нм, окруженные остаточной аморфной фазой (30 %) [6].

Добавление того или иного элемента повлечет за собой изменения в процессе кристаллизации. Это повлияет на структуру сплава, чем будет обусловлены изменения свойств, в том числе и магнитных.

Помимо работ, посвященных изменению свойств сплавов при вариации соотношения химических элементов классического состава FINEMET [7-12], есть также исследования о измененных в большей или меньшей степени составов нанокристаллических сплавов [13-19] с добавлением других химических элементов, например Y, P, Ga, Al, Ge и т.д.

В работе [13] при добавлении Ge, а в работе [18] при добавлении Mo и V, улучшалась термостабильность сплавов. В результате замещения ниобия – элемента, сдерживающего рост зерна, в работе [19] происходило изменение магнитных свойств за счет изменения размера зерна.

Для модификации кристаллической структуры и как следствие свойств сплава варьируется не только химических состав, но и применяются различные методы обработки. Так, например, при изменении температуры и времени отжига сплава одинакового состава [20-23] его структура, размер зерен повлекли к изменению коэффициента магнитострикции, коэрцитивной силы и т.д.

Другой метод обработки – это воздействие магнитного поля на образец во время отжига. В работе [21] показано, что при воздействии поперечного магнитного поля при отжиге магнитная проницаемость значительно увеличилась, а потери на перемагничивание, наоборот, уменьшились. В статье [24] авторы продемонстрировали увеличение магнитной индукции насыщения классического FINEMET за счет приложения продольного магнитного поля. Применение магнитного поля во время отжига создает наведенную магнитную

анизотропию, что приводит к изменению свойств без варьирования состава и температуры отжига.

Работы [21, 25, 26] содержат выводы о существовании зависимости магнитных свойств от толщины лент и пленок.

На сегодняшний день существует большой объем разнообразных исследовательских работ с представленными в них экспериментальными данными. При наличии большого объема данных в этой области применение методов машинного обучения может произвести значительный вклад в развитие существующих и создание новых модификаций нанокристаллических сплавов, что позволит исправить недостатки существующих материалов.

1.2 Машинное обучение

Машинное обучение заключается в извлечении знаний из большого массива данных. Машинное обучение находится на пересечении статистики, искусственного интеллекта и информационных технологий. Специализация заключается в использовании данных и алгоритмов для имитации процесса наработки опыта человеком с постепенным повышением точности [27].

Машинное обучение является важным инструментом в процессе формирования выводов из массивных наборов данных.

Цель машинного обучения - решение сложных профессиональных задач на основе массива данных, систематизация которых сложна для человека.

Существуют принципиальные различия между разработкой физических моделей и моделей машинного обучения. При традиционной разработке используются входные данные и алгоритм (функция зависимости выходных данных от входных). В результате работы физической модели будут получены выходные данные.

Машинное обучение в свою очередь использует входные и выходные данные для формирования неизвестной функции зависимости между этими данными. Дальнейшая экстраполяция рассчитанной функции позволяет делать прогноз для принятия решений.

Существует три составляющие машинного обучения:

Алгоритм. Каждую задачу можно решить разными способами. Для разных целей можно подобрать разные алгоритмы. Суть алгоритмов машинного обучения сводится к одному: по данным найти неизвестную зависимость, а по найденной зависимости предсказать результат.

Данные. Данные являются важным фактором качества предсказаний. Чем качественнее данные (т. е. более полные, разнообразные, многочисленные), тем эффективнее

обучение. Сбору и составлению баз данных при создании моделей обучения уделяют особое внимание. В последствии данные делятся на обучающую и тестовую выборку. Обучающие данные нужны для построения аналитической зависимости, а тестовые – для контроля универсальности этой зависимости.

Признаки. Это набор свойств, характеристик или признаков, которые описывают модель. Признаками являются характеристики, косвенно или прямо влияющие на целевую характеристику – данные, которые необходимо спрогнозировать.

Например, температура отжига – это признак, который влияет на целевую характеристику - коэрцитивную силу.

1.3 Жизненный цикл модели машинного обучения

Для выполнения задач анализа данных специалисты по машинному обучению часто используют международный стандарт CRISP-DM, представленный в 1999 году, который включает в себя все стадии жизненного цикла анализа данных, представленные на рис. 1.1.

CRoss Industry Standard Process for Data Mining (CRISP-DM) – стандарт, описывающий общие процессы и подходы к аналитике данных, используемые в промышленных проектах независимо от конкретной задачи и индустрии.

Рисунок 1.1 - Жизненный цикл исследования данных по CRISP-DM [28]

По аналогии с межгосударственными стандартами ГОСТ использование CRISP-DM в исследовании позволяет достигнуть унификации, а также упростить планирование исследования.

Жизненный цикл модели машинного обучения включает следующие этапы:

1 Определение целей исследования и формирования требований. На данном этапе составляется план проекта, назначаются исполнители и ответственные лица, а также формируется основная цель исследования. Этот этап непосредственно связан со следующими, поскольку весь процесс работы должен быть направлен на достижение определенной цели.

2 Анализ данных. После сбора информации из источников, которые соответствуют цели проекта, следует этап анализа этих данных. Необходимо выявить проблемы собранных данных, понять какие сведения доступны, а также сформировать гипотезы о наличии закономерностей.

3 Выявленные на предыдущем этапе проблемы, необходимо исправить. На этом этапе получают итоговый набор данных, который будет поступать непосредственно на вход модели.

4 На этом шаге проекта применяются различные алгоритмы и методики моделирования. За счет разных принципов работы этих алгоритмов вид данных может требовать существенных изменений, поэтому часто необходим возврат на предыдущий этап, с целью изменения диапазонов, типа распределения или другой обработки данных.

5 После завершения моделирования, происходит оценка существующей модели машинного обучения. По результатам анализа характеристик модели можно понять, какие из задач по плану, составленному на первом этапе, удалось достичь, а для каких необходима дополнительная работа. По результатам этого шага составляется отчет о проделанной работе, а также определяются шаги для последующей работы.

6 Внедрение в последующую работу является логическим завершающих шагом в процессе разработки модели машинного обучения. После внедрения происходит плановый мониторинг, а также поддержка разработанной модели, в том числе с помощью новых моделей, новых данных, которые могут быть получены на следующем цикле исследования.

Таким образом, жизненный цикл модели замыкается в привычный нескончаемый процесс постоянного поддержания и совершенствования продукта.

1.4 Валидация

В 2020 году были впервые введены национальные стандарты [29 - 31], посвященные контролю качества средств, технологий или их совокупности, предназначенных для автоматизированного анализа данных в области мониторинга и прогнозирования поведения людей [29], распознавания при досмотре незаконных вложений по теневым рентгеновским изображениям [30] и распознавания голосовых команд управления [31].

Эти документы представляют собой методики определения показателей качества, устанавливающие требования к валидационным данным и процессу расчета показателей качества.

Перечисленные стандарты относятся к задачам классификации, решаемые с помощью машинного обучения, однако на сегодняшний день не существует регламентирующих документов для задач регрессии ни в частном, ни в общем виде.

Более того, стандарты [29-31] не устанавливают требования к алгоритмам, на основе которых работают модели анализа. В ходе данной магистерской диссертации будет показана важность соблюдения требований, например, отсутствие переобучения, в процессе анализа данных. Контроль работы модели анализа данных на основе машинного обучения наиболее полно может быть выполнен в результате ее валидации.

Согласно межгосударственному стандарту словарю информационных технологий [32] под валидацией понимается подтверждение путем проверки и предоставления объективных доказательств выполнения особых требований к конкретному предусмотренному применению, а также того, что все требования выполняются надлежащим образом и в полном объеме, и что обеспечивается прослеживание выполнения системных требований.

Валидация моделей машинного обучения – это комплексный процесс, который должен быть внедрен во весь жизненный цикл исследования. На рис. 1.2 показан жизненный цикл модели машинного обучения, на котором отмечена интеграция валидационного процесса в этапы разработки и эксплуатации модели.

Рисунок 1.2 – Валидация модели машинного обучения на разных этапах жизненного цикла модели [33]

1.4.1 Валидация данных

Первый этап валидации модели наступает на этапе анализа данных. Поскольку данные являются основополагающей частью модели машинного обучения, от которой зависит результат прогноза, в науке о данных (от англ. Data Science) существует процесс валидации или профилирования данных.

Поскольку информационные технологии – это продукт совместных разработок международного сообщества, существуют различные термины: валидация и профилирование (от англ. Profile) данных – которые, по существу, означают процесс подтверждения с помощью качественных и количественных показателей соответствия требованиям качества данных, требованиям цели исследования, установленных в результате постановки задачи.

В соответствии со стандартом ISO/IEC 25012:2008 и [34] основными критериями качества данных являются точность, достоверность, полнота и согласованность. В ходе валидации данных выявляют проблемы данных, которые описываются в отчете о валидации модели. Некоторые из проблем, выявленные на этом этапе, являются критическими для работы модели и ее результатов, поэтому по результатам валидации данных применяются различные методы для решения выявленных проблем (например, восстановление пропусков, исключение аномалий и т.д.)

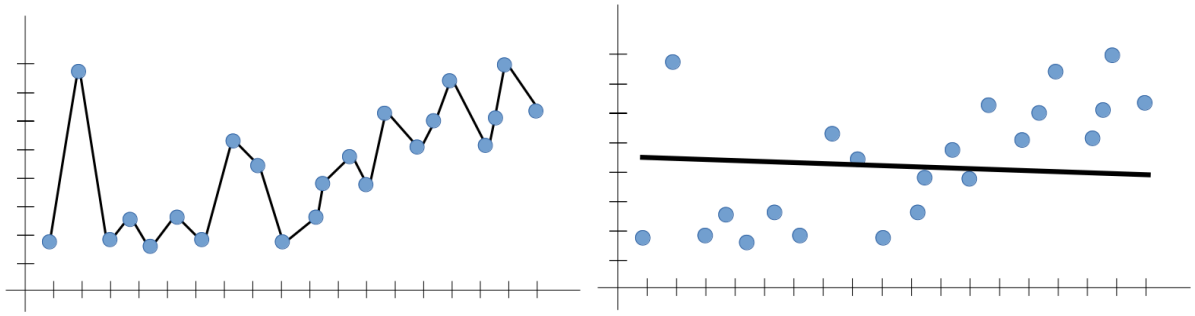
1.4.2 Первичная валидация

Как обозначалось ранее не существует универсального алгоритма. Выбор подходящего алгоритма и оптимальных гиперпараметров – задача для разработчика на этапе разработки модели.

С точки зрения валидации необходимо проконтролировать правильность работы модели, посредством оценки предсказательной способности и точности, а также проверки на обобщающую способность.

Обобщающая способность — это способность аналитической модели, построенной с применением алгоритмов машинного обучения выдавать правильные результаты не только для данных обучения (на которых строилась аналитическая модель), но и для новых неизвестных для модели данных.

Считается, что модель обладает обобщающей способностью, когда погрешность прогноза на тестовых данных достаточно мала (или предсказуема) и не сильно отличается от погрешности прогноза на обучающих данных. Крайние положения обобщающей способности продемонстрированы на рис. 1.3 – это переобучение и недообучение.



а) б)

Рисунок 1.3 - Обобщающая способность:

а - переобучение, б – недообучение.

Переобучение (рис. 1.3, а) возникает в результате слишком сложных моделей зависимости. В таком случае погрешность на обучающих данных существенно ниже, чем на тестовых.

Недообучение (рис. 1.3, б) возникает в результате слишком простых моделей зависимости.

Для контроля обучающей способности применяется разделение данных и перекрестная проверка (от англ. cross-validation).

Тестовая и тренировочные выборки

Обучение и проверка модели невозможна по одним и тем же данным.

В машинном обучении данные разделяют на 3 составляющие – рис. 1.4:

Данные обучения (Training data) — не менее 60% данных должно использоваться для обучения.

Данные тестирования (Test data) — этот набор данных используется для тестирования модели после её полного обучения. Данные в наборе тестирования должны выглядеть точно так же, как будут выглядеть реальные данные после развёртывания модели.

Данные валидации (Validation data) — выборка (10–20%) из общего набора данных. С помощью этих данных производится оценка и контроль модели во время обучения.

Главное в формировании выборок – не объединять обучающие данные с оценочными (тестовым и валидационным), поскольку это грозит переобучением модели. В этом случае модель получит высокие оценки качества в процессе тренировки, но не покажет такого результата на реальных данных.

Рисунок 1.4 - Разделение данных

Кросс-валидация

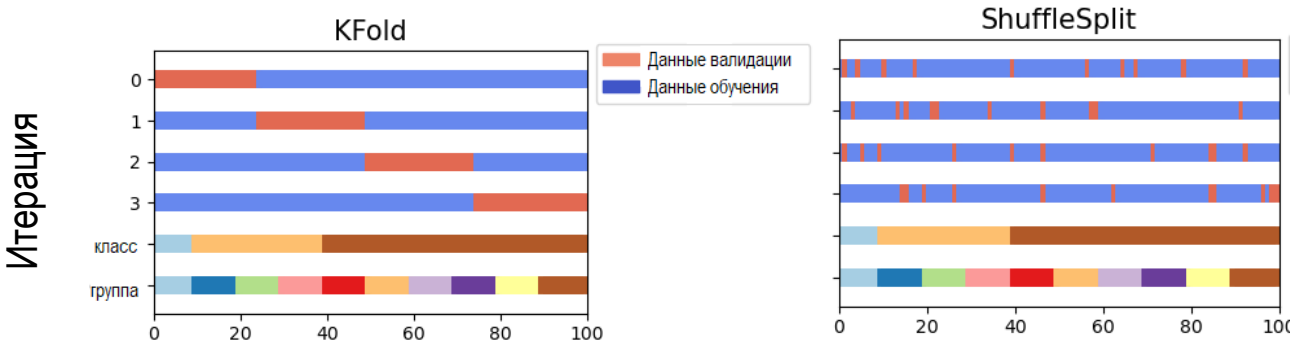
При кросс-валидации, обучающие данные делятся на равные части. Алгоритм обучается, используя все, кроме одной из частей, а тестируется - на оставшейся. Части могут затем меняться несколько раз так, что алгоритм обучается и оценивается на всех данных. В таблице 1.1 показан пример перекрестной проверки с данными, разбитыми на пять частей.

Таблица 1.1 - Разбиение данных при кросс-валидации

	A	*B*	*C*	*D*	*E*
1 итерация	Данные валидации	Обучающие данные	Обучающие данные	Обучающие данные	Обучающие данные
2 итерация	Обучающие данные	Данные валидации	Обучающие данные	Обучающие данные	Обучающие данные
3 итерация	Обучающие данные	Обучающие данные	Данные валидации	Обучающие данные	Обучающие данные
4 итерация	Обучающие данные	Обучающие данные	Обучающие данные	Данные валидации	Обучающие данные
5 итерация	Обучающие данные	Обучающие данные	Обучающие данные	Обучающие данные	Данные валидации

Кросс-валидация дает более точную оценку эффективности модели, чем тестирование с использованием только одной части данных.

Но простое разбиение, как показано на рис. 1.5, а и табл. 1.1 может привести к точному описанию одного диапазона свойства, но некорректному описанию другого, за счет неоднородности распределения наших данных. Поэтому в работе мы использовали метод выделения проверочных данных вперемешку (shuffled) рис.1.5, б, который позволяет создавать проверочный набор данных одинакового размера, но из разных диапазонов значений магнитных свойств.



а) б)

Рисунок 1.5 - Разные виды разбиения данных при кросс-валидации:

а – простое разбиение, б – разбиение вперемешку [35].

2 Постановка задачи

Исходя из приведенного анализа литературы, можно отметить актуальность применения моделей машинного обучения в области разработки магнитных материалов.

Несмотря на ажиотаж проведения исследований в различных областях с использованием машинного обучения, развитие метрологического обеспечения этой цифровой технологии не происходит, что может быть обусловлено новизной технологии.

В данной работе будет присутствовать как процесс разработки модели машинного обучения, так и процесс ее валидации, который внедрен в различные этапы исследования.

Поэтому цель настоящей работы не ограничивается разработкой модели. Исходя из описанного выше, были сформулированы следующие цели магистерской диссертации:

1. разработка модели машинного обучения для прогноза магнитных характеристик нанокристаллического сплава типа FINEMET на основе химического состава и термической обработки;
2. валидация разработанной модели машинного обучения на разных этапах жизненного цикла.

Для достижения указанных целей были поставлены и решались следующие задачи:

1) изучение библиотек для разработки модели машинного обучения на языке программирования Python с открытым исходным кодом;
2) сбор данных для создания и валидации модели;
3) создание пакета программ для обработки данных;
4) выбор алгоритмов;
5) подбор оптимальных гиперпараметров выбранных алгоритмов;
6) обучение модели и визуализация полученных результатов обучения;
7) верификация полученных результатов.

3 Методика

Основная методика выполнения работы – это компьютерный эксперимент, направленный на прогнозирование магнитных свойств, основываясь на составе и термической обработки сплава.

Созданный пакет программ выполнял различные функции:

- визуализация и обработка данных;
- поиск оптимальных гиперпараметров с использованием библиотеки Optuna [36];
- непосредственно сама модель машинного обучения, написанная в среде разработки Jupyter Notebook.

Jupyter Notebook – это среда разработки, в которой имеется возможность разделения программного кода на отдельные блоки для воспроизведения их в произвольном порядке. Помимо этого, преимущество этой среды разработки заключается в возможности вывода результатов блока (графиков, таблиц, расчетов) непосредственно после его выполнения. Более того, внутри программы доступно написание инструкций к каждому блоку, что упрощает понимание функционала блока программ, делает доступным совместную работу над проектом, а также делает удобнее процесс валидации на разных этапах жизненного цикла модели.

3.1 Сбор данных

Для решения поставленной задачи источниками данных должны быть академическая литература и патентная информация из опубликованных экспериментальных работ. Работы должны описывать нанокристаллические сплавы FINEMET и их магнитные характеристики.

Набор данных является одной из главных составляющих машинного обучения. Точность модели и предсказанных результатов напрямую зависит от их качества.

Для обучения модели машинного обучения была использована база данных, представленная в статье [2] Wang и соавторов. В нее вошли статьи об исследованиях нанокристаллических сплавах FINEMET, опубликованные в открытых источниках с 1988 по 2018 годы.

Для извлечения данных использовалось несколько способов:

- Парсинг – программное извлечение из текста необходимой информации. Источником для программного извлечения стала открытая база SCOPUS. Поскольку в текстах могут быть расхождения в обозначениях, единицах измерения или попросту неожиданный

для программы способ описания характеристики, извлечение данных может быть затруднительным. Поэтому применяют способ сбора:

- вручную, путем просмотра статьи и занесения сведений в базу данных. Но даже здесь можно столкнуться с трудностями, когда значения физических величин указаны не в тексте статьи, а на графиках. В таком случае можно использовать бесплатную программу WebPlotDigitizer, которая позволяет оцифровывать графики.

После окончания сбора данных в единую базу необходимо провести анализ и описание данных.

3.2 Анализ данных

Анализ данных является одним из методов, позволяющих исключить поступление на вход информационной системы или её компонент заведомо ошибочных, неполных или неточных данных, которые могут привести к ошибочным результатам работы. В процессе анализа могут осуществляться корректировка либо исключение данных, файлов, пакетов и записей. В результате анализа данных формируется пункт в отчете о валидации, включающий в себя анализ данных с помощью количественных и качественных параметров, соответствие данных поставленной задаче, а также описание особенностей данных. С помощью анализа данных принимается решение по выбору алгоритма машинного обучения для его реализации в модели, а также метрик по оценке качества работы модели.

База данных насчитывает в себе информацию о 1440 образцах.

Произведем поиск проблем:

- *Проверка типа данных*: Все данные, представленные для анализа, соответствуют необходимым типам данных.

- *Проверка формата данных* (единства единиц измерения и их соответствие физической величине): физические величины имеют одинаковые и согласованные единицы измерения.

- *Проверка недопустимых значений* (например сумма хим. элементов превышает 100, отрицательное содержание хим. элемента): все значения данных допустимы.

- *Проверка диапазона значений физических величин:* обнаружено несоответствие диапазона значений размера зерна и температуры отжига цели исследования.

Таблица 3.1 - Диапазоны некоторых физических величин в собранных данных

Параметр	Диапазон
Fe	(63,5 – 92,5) ат.%
Au	(0 – 1) ат.%
Температура отжига	(79 – 1174) К
Время отжига	(60 – 951000) с
Размер зерна	(2 – 298) нм
Коэрцитивная сила	(0,022 – 10149) А/м
Индукция насыщения	(0,019 – 1,94) Тл
Магнитная проницаемость	(16 – 500748)

В таблице 3.1 представлены примеры диапазонов признаков и целевых характеристик, в том числе и тех, где присутствует несоответствие.

Поскольку цель исследования - сплавы нанокристаллической структуры, ограничим размер зерна 60 нм. Под отжигом понимается процесс нагрева сплава в течение некоторого времени, а затем медленного охлаждения. Превращения при температурах ниже 323 К, происходят при особых условиях, которые не учитываются в будущей модели машинного обучения, поэтому ограничим диапазон значений температуры отжига (323 – 1174) К.

При текущих ограничениях в базе содержится информация о 1294 образцах.

- *Проверка пропущенных значений:*

Химический состав (с элементами Fe, Si, C, Al, B, P, Ga, Ge, Cu, Ag, Au, Zn, Ti, V, Cr, Zr, Nb, Mo, Hf, Ta, W, Ce, Pr, Gd, U) и условия отжига (температура, время, приложенное магнитное поле) указаны для всех объектов.

В таблице 3.2 приведены результаты проверки пропусков данных после введения ограничений, указанных в предыдущем пункте. Стоит обратить внимание, что количество данных для целевых переменных «суммарные потери» и «электрическое сопротивление» меньше 50 образцов, а для температуры Кюри 94. Такое малое значение данных может оказать влияние на работу и результат модели машинного обучения.

Таблица 3.2 – Результат проверки пропусков в данных после очистки

Признаки	Кол-во объектов без пропусков
Хим.элементы	1294
Условия отжига	1294
Температура начала первичной кристаллизации (К)	115
Температура пика первичной кристаллизации (К)	529
Температура пика вторичной кристаллизации (К)	454
Толщина ленты (нм)	1028

Продолжение таблицы 3.2

Целевые характеристики	Кол-во объектов без пропусков
Коэрцитивная сила (А/м)	741
Температура Кюри (К)	94
Суммарные потери (кДж/м^3 при В = 0.2 Тл 100 кГц)	38
Электрическое сопротивление (10^{-6} х Ом х м)	48
Магнитная проницаемость (при 1 кГц)	371
Коэф. магнитострикции (10^{-6})	203
Магнитная индукция насыщения (Тл)	294
Размер зерна (нм)	204

- *Проверка корреляций и взаимосвязей*

Построим корреляционные матрицы для целевых характеристик и признаков. Число на матрицах рис. 3.1, 3.2, 3.3 – коэффициент линейной корреляции Пирсона, который позволяет оценить взаимосвязь двух переменных. Диапазон значений коэффициента корреляции от -1 до +1. В случае независимости двух переменных коэффициент корреляции имеет значение 0. Отсутствие коэффициента корреляции в матрице означает отсутствие соответствующих данных.

Корреляции целевых характеристик и химических элементов представлена на рис. 3.1. Наибольшие значения коэффициента корреляции с целевыми характеристиками имеют Fe, Si, Cu, Nb – эти элементы входят в классический состав FINEMET и именно благодаря этих элементам FINEMET имеет отличные магнитомягкие свойства.

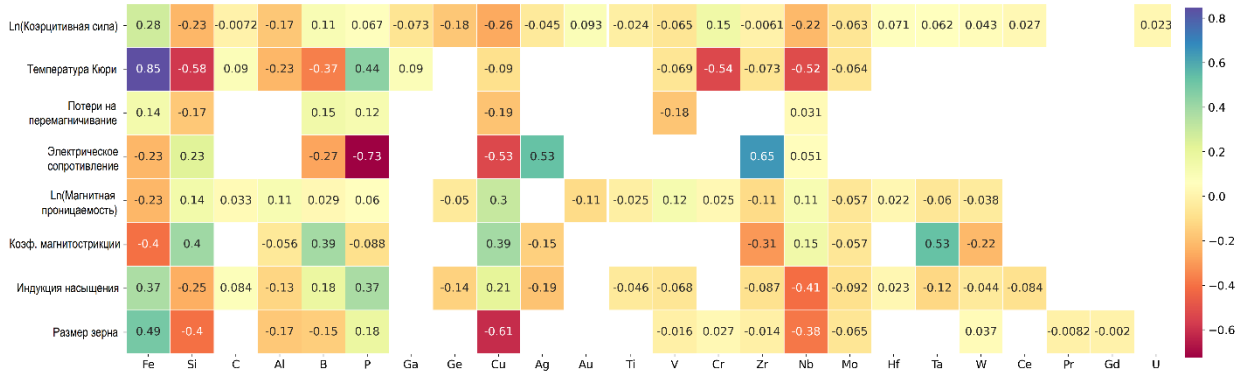


Рисунок 3.1 – Матрица корреляций целевых характеристик и химических элементов

На рис. 3.2 представлены корреляции целевых характеристик и условий термообработки. В данной группе признаков корреляции с магнитными свойствами выражены значительнее. Однако ни один из признаков не имеет коэффициентов корреляции выше 0,85.

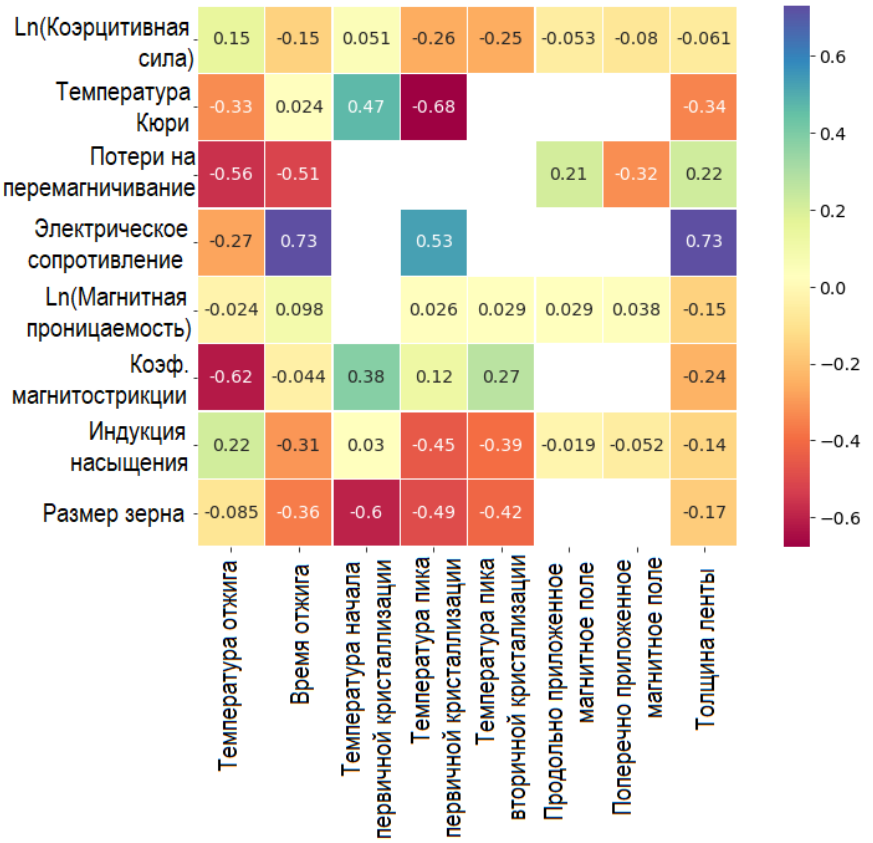


Рисунок 3.2 – Матрица корреляций целевых характеристик и условий обработки

Стоит отметить почти нулевую корреляцию магнитной проницаемости и коэрцитивной силы от имеющихся признаков. В дальнейшем это может затруднять работу модели и влиять на окончательный результат.

Рассмотрим также корреляцию целевых характеристик между собой. Как видно из рис. 3.3 некоторые целевые характеристики сильно коррелированы. Особенно высоки значения коэффициентов корреляции для потерь на перемагничивание, электрического сопротивления и температуры Кюри. Для этих характеристик рассмотрим линии регрессии на рис. 3.4. Из представленного рисунка видно, что корреляции этих характеристик значительны за счет малого количества данных.

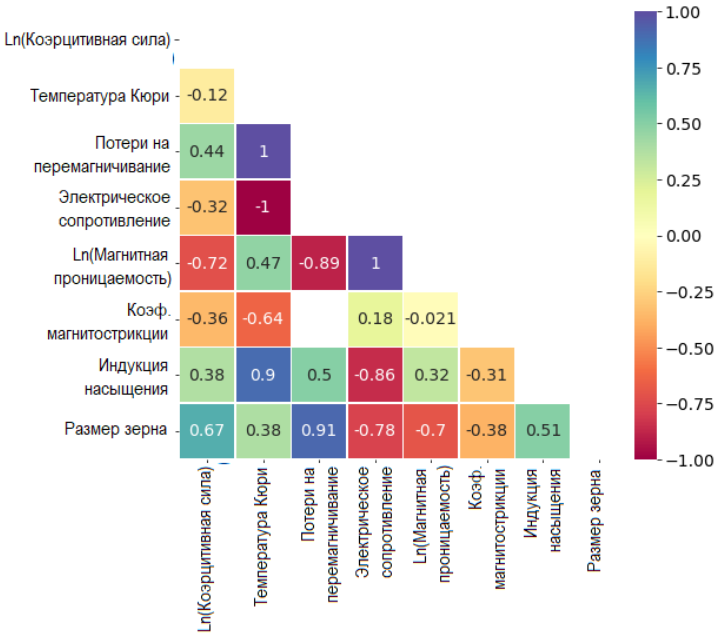


Рисунок 3.3 – Матрица корреляций целевых характеристик между собой

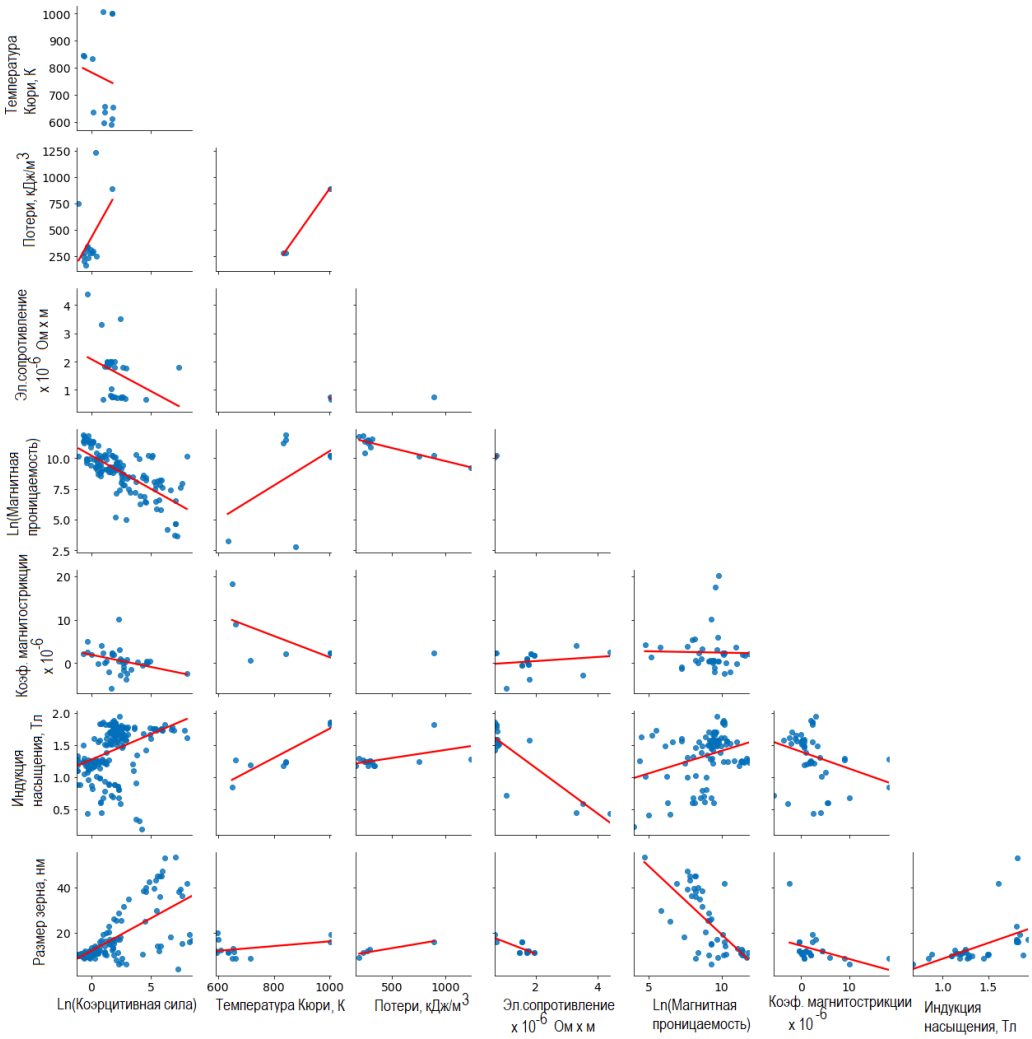
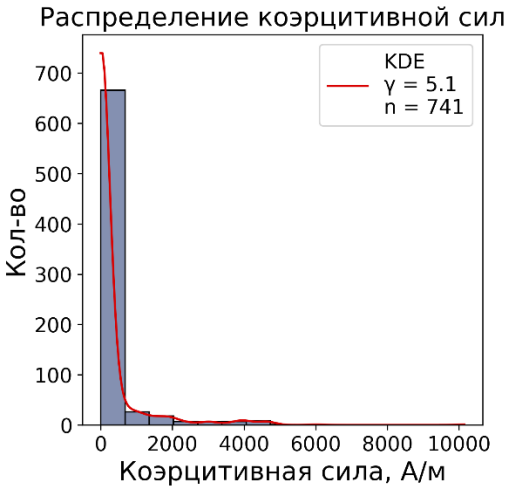


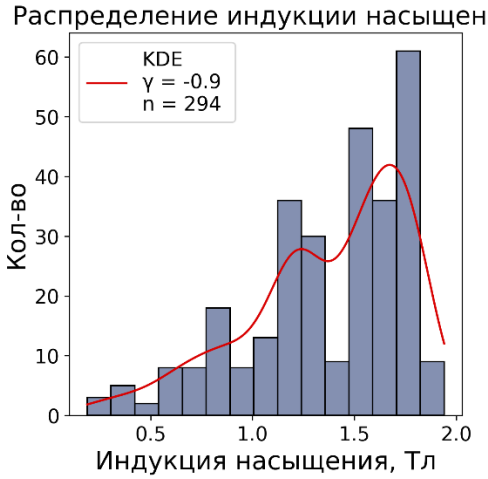
Рисунок 3.4 – Распределение данных в координатах целевых характеристик с линиями регрессии

- *Проверка распределения данных:*

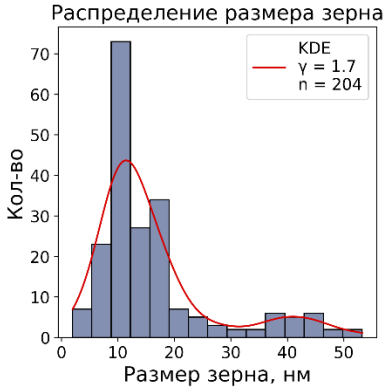
Распределение целевых характеристик:



а)

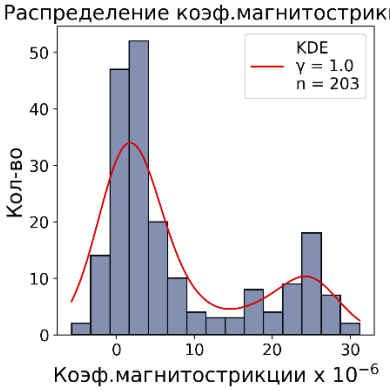


б)

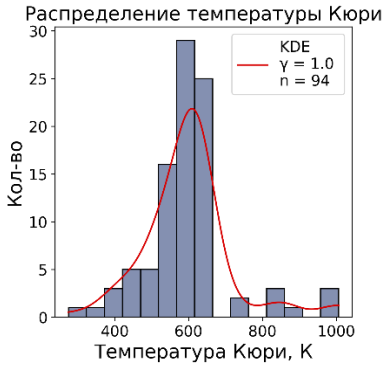


в)

г)

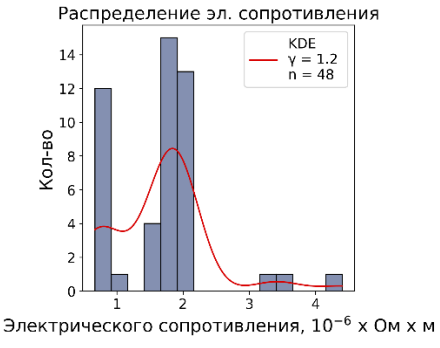


д)



е)

ж)



з)

Рисунок 3.5 - Распределения целевых характеристик до обработки: а – коэрцитивной силы, б – индукции насыщения, в – размера зерна, г – магнитной проницаемости, д – коэффициента магнитострикции, е – температуры Кюри, ж – потерь сердечника на перемагничивание, з – электрического сопротивления, γ – перекос данных, n – количество образцов с указанной характеристикой

Распределения химических элементов в составах образцов:

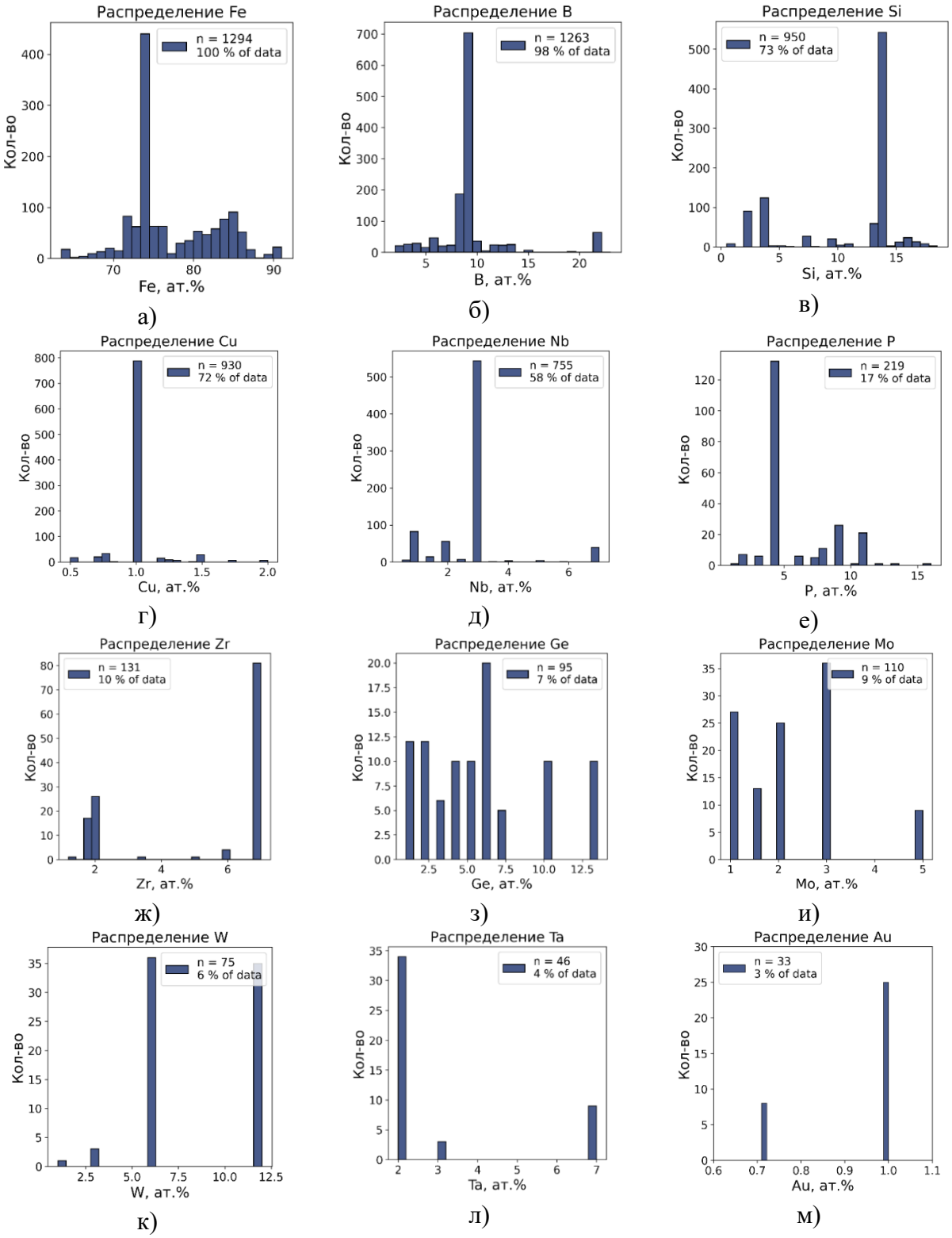


Рисунок 3.6 - Распределения химических элементов: а – Fe железо, б – B бор, в – Si кремний, г – Cu медь, д – Nb ниобий, е – P фтор, ж – Zr цирконий, з – Ge германий, и – Mo молибден, к – W вольфрам, л – Ta тантал, м – Au золото, n – количество образцов с ненулевым значением

Выводы:

На рис. 3.5, а и г показано распределение коэрцитивной силы и магнитной проницаемости, графики которых имеют значительные перекосы. Большая часть данных находится в диапазоне от 0 до 2000 А/м – для коэрцитивной силы, от 0 до 100000 – для магнитной проницаемости. Для удобства анализа воспользуемся логарифмированием значений.

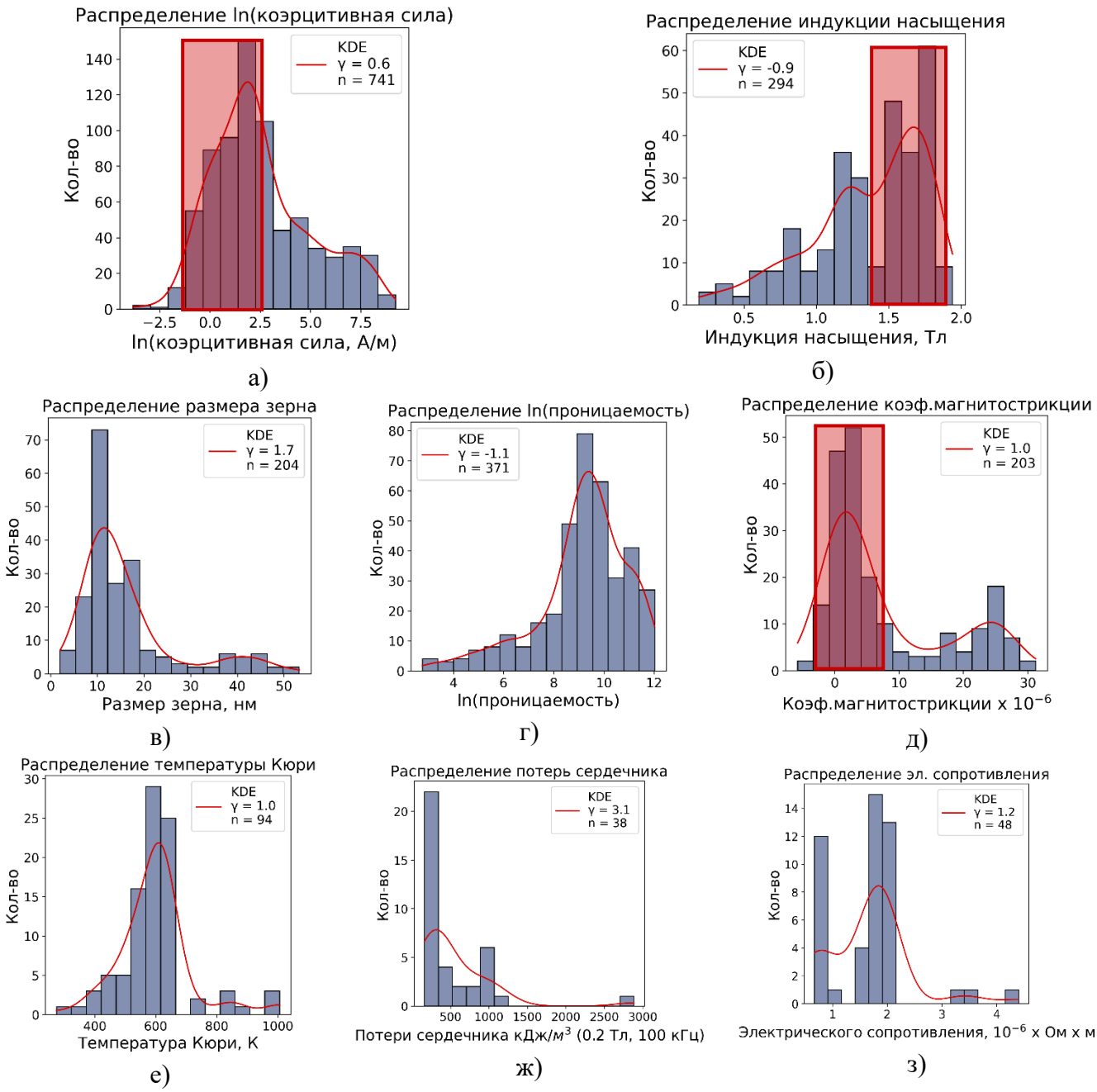


Рисунок 3.7 – Распределения целевых характеристик после обработки: а – натурального логарифма коэрцитивной силы, б – индукции насыщения, в – размера зерна, г – натурального логарифма магнитной проницаемости, д – коэффициента магнитострикции, е – температуры Кюри, ж – потерь сердечника на перемагничивание, з – электрического сопротивления, γ – перекос данных, n – количество образцов с указанной характеристикой

В распределениях магнитных характеристик образцов видны несколько проблем набора данных.

1) При исследовании материалов авторы статей предпочитают публиковать уникальные/рекордные/привлекательные результаты (т.е. большую индукцию насыщения, малую коэрцитивную силу и т. д.) – красные прямоугольники на рис. 3.7. Такие данные составляют около 45 %. В результате чего в распределении данных появляются перекосы и двойные максимумы.

2) Отсутствие всех магнитных свойств для всех образцов. Причиной этому может быть сложность измерения той или иной величины.

3) Распределение Fe на рис. 3.6, а показывает, что преобладающий состав - 73,5 ат.%, что является исходным атомным процентным содержанием Fe в FINEMET.
Также по распределениям на рис. 3.6 можно отметить присутствие «модных» составов для изучения. Поэтому разнообразное варьирование химических элементов ограничено.

4) Данные не подходят для анализа с помощью алгоритмов, работающих на основе учета расстояния между точками, поскольку расстояния между точками разных признаков различное. Об этом свидетельствует различие диапазонов значений, представленных в таблице 3.1.

Стратегии по устранению проблем:

Проблемы 1 – 3 невозможно решить с помощью обработки. Это особенность данных, присущая экспериментальным источникам. Изменить ситуацию может лишь дополнительный поиск данных.

Проблема 4 устранима с помощью преобразований.

При реализации регрессионной модели, признак, изменяющийся в диапазоне от -1 до 1, имеет меньшее влияние на обучение, чем признак, имеющий диапазон от 1 до 1000, несмотря на важность признака в целом.

Как видно из таблицы 3.1 и рисунков 3.5, 3.6, 3.7 с распределениями величин, данные распределены ненормально, в различных диапазонах, с перекосами. Для корректной работы алгоритмов проведем нормализацию и стандартизацию данных.

Подготовка данных:

Нормализация и стандартизация – методы предобработки признаков, приведение их к общей шкале без потери информации.

Нормализация меняет диапазон данных без искажения формы распределения, а стандартизация меняет форму распределения, приводя ее к нормальному.

Для стандартизации данных воспользовались самым распространенным методом Z стандартизацией. Данные преобразуются по формуле:

$$z = \frac{x - \mu}{\sigma} \tag{3.6}$$

где x – признак i-ого объекта, μ - среднее значение по признаку, σ – стандартное отклонение по признаку.

В результате стандартизации признак имеет среднее значение $\mu = 0$ и ско $\sigma = 1$.

Для масштабирования использовался метод MinMaxScaler, который переводит каждый признак в диапазон от 0 до 1 по формуле:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.2}$$

где x_{scaled} – масштабированный признак i-ого объекта, x – признак i-ого объекта, x_{min} – минимальное значение признака, x_{max} – максимальное значение признака.

3.3 Отбор важных признаков

Некоторые признаки в данных при построении модели машинного обучения являются важными, а другие являются шумовыми. Удаление избыточных признаков может улучшить точность, легкость интерпретации и производительность модели.

Для определения важности признаков используем метод главных компонент (от англ. principal component analysis PCA) и алгоритм случайный лес (от англ. Random Forest).

Отбор признаков выполняется во время обучения модели, оптимизируя их набор для достижения лучшей точности.

На рис. 3.8 приведен пример ранжирования важных признаков по величине нормированной важности для прогноза коэрцитивной силы, полученный с помощью алгоритма Random Forest.

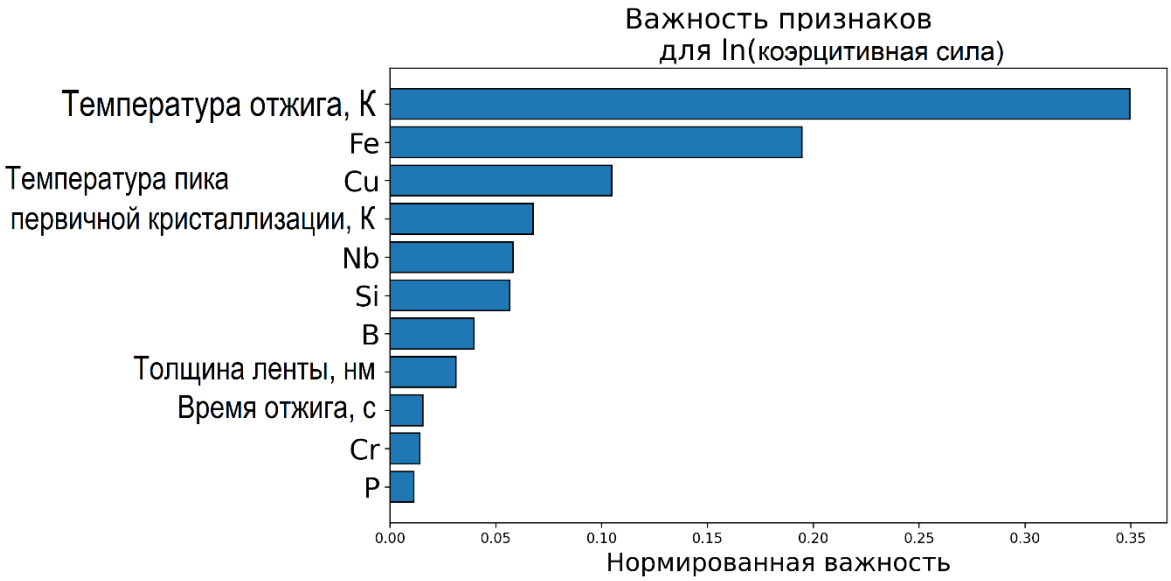


Рисунок 3.8 - Важность признаков для логарифма коэрцитивной силы

Отбор важных признаков происходит по пределу значения совокупной важности признака. На рис. 3.9 показана кривая совокупной важности от количества признаков. После определения предела числа признаков (предел показан на рис. 3.9 синей пунктирной линией) значение совокупной важности выходит на плато. Установим предел числа признаков для достижения совокупной важности 0,95.

Помимо этого, удалим признаки, в которых пропущено более 50 % данных.

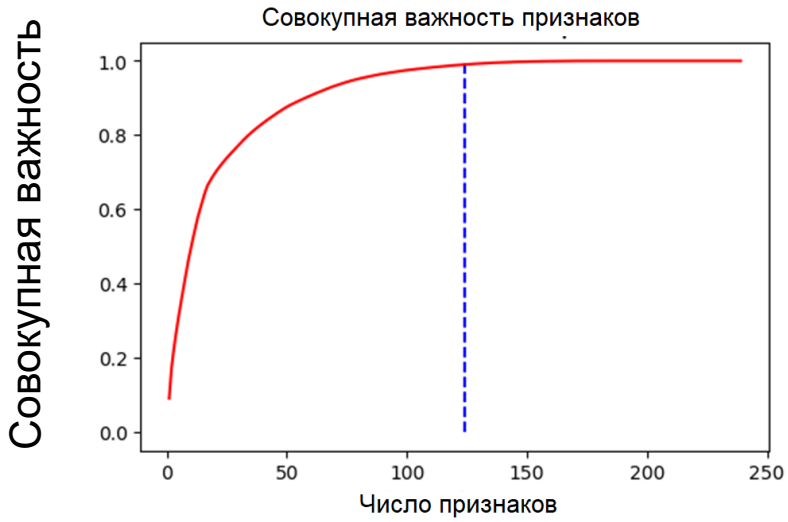


Рисунок 3.9 - График важности от количества признаков

В итоге имеем набор важных признаков для каждой целевой характеристики, представленный в таблице 3.3.

Таблица 3.3 – Наборы важных признаков для целевых характеристик

Целевая характеристика	Хим. элементы	Условия обработки
Коэрцитивная сила	Fe Si B P Ge Cu Au Nb Cr Mo Zr Ce	Температура отжига, К Температура пика первичной кристаллизации, К Толщина ленты, нм Время отжига, с
Размер зерна	Fe Si B Cu Nb P Cr	Температура отжига, К Толщина ленты, нм
Индукция насыщения	Fe Si B P Cu Nb	Температура отжига, К Температура пика первичной кристаллизации, К Толщина ленты, нм Время отжига, с Температура пика вторичной кристаллизации, К
Магнитострикция	Fe Si B Nb Ta	Температура отжига, К Толщина ленты, нм Время отжига, с
Магнитная проницаемость	Fe Si B Nb Ta Cu Cr Au	Температура отжига, К Толщина ленты, нм Время отжига, с
Температура Кюри	Fe Si B Nb P Cr	Температура отжига, К Время отжига, с Температура начала первичной кристаллизации, К

Продолжение таблицы 3.3

Электрическое сопротивление	Fe Si Cu P Ag	Температура отжига, К Толщина ленты, нм Время отжига, с
Потери на перемагничивание	Fe	Температура отжига, К Время отжига, с Продольное магнитное поле при отжиге Толщина ленты, нм

3.4 Выбор метрики оценки алгоритма

Оценка качества предсказаний играет важную роль в выборе алгоритмов и их параметров, поэтому необходимо понимать, какие существуют метрики качества и какие метрики подходят для текущей задачи.

Поскольку мы решаем задачу регрессии с помощью машинного обучения для оценки такого типа задач используются следующие метрики качества:

MSE - Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2 , \quad (3.3)$$

где n – кол-во значений, $a(x_i)$ – предсказанное значение, y_i – реальное значение.

Чем больше *MSE*, тем хуже модель.

MSE применяется в ситуациях, когда нам надо подчеркнуть большую разницу предсказанных и реальных значений. Исходя из этого, мы можем выбрать модель, которая дает малое количество больших погрешностей прогноза. Значительные погрешности прогноза становятся заметнее за счет того, что разница значений возводится в квадрат.

Также за счет своей дифференцируемости эффективно используется для поиска минимальных и максимальных значений с помощью математических методов.

Стоит отметить, что оценку *MSE* тяжело интерпретировать. Неясно хороша модель или плоха при *MSE* = 5. Однако, *MSE* может быть использована на этапе обучения для минимизации разброса данных.

RMSE – Root Mean Squared Error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2}\,. \tag{3.4}$$

RMSE – это квадратный корень из *MSE*. Эта метрика так же, как и *MSE* чувствительна к выбросам.

В отличие от *MSE, RSME* лучше подходит для интерпретации, поскольку выражается в единицах измерения исходных данных, но не используется при сравнении предсказаний различных по единицам величин данных.

MAE - Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - y_i|. \tag{3.5}$$

Эта метрика похожа на *MSE*, однако у нее есть существенное отличие – она менее чувствительна к выбросам (из-за отсутствия возведения разницы в квадрат). Минимизация *MAE* используется для достижения правдоподобия модели, как и *MSE*.

Так или иначе, метрики, описанные выше, не позволяют нам оценить модели между собой.

R^2 – коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.6}$$

R^2 является показателем качества регрессионной модели и принимает значения от 0 до 1. Фактически, данная мера качества — это нормированная *MSE*. Значение 1 соответствует идеальной прогнозирующей способности, а значение 0 соответствует константе модели, которая лишь предсказывает среднее значение ответов в обучающем наборе [37].

Коэффициент детерминации подходит для сравнения нескольких разных величин, поскольку не зависит от единиц измерения данных.

Моделирование:

3.5 Выбор алгоритмов

Важно понимать, не существует плохих и хороших алгоритмов. Каждый алгоритм уникален и подходит при определенных условиях. Работа алгоритма существенно зависит от входных данных.

Существуют различные группы алгоритмов, используемые для задач регрессии в машинном обучении:

Линейные,

Нелинейные,

Ансамбли,

Нейронные сети

Выбор нужного алгоритма – является важной задачей при создании модели машинного обучения.

Первым этапом выбора алгоритмов стала блиц проверка алгоритмов:

Линейные алгоритмы:

— Линейная регрессия / Linear Regression (‘LR’)

— Гребневая регрессия (ридж-регрессия) / Ridge Regression (‘R’)

— Лассо-регрессия (от англ. LASSO — Least Absolute Shrinkage and Selection Operator) / Lasso Regression (‘L’)

— Метод регрессии «Эластичная сеть» / Elastic Net Regression (‘ELN’)

— Байесовская гребневая регрессия / Bayesian ridge regression (‘BR’)

Нелинейные алгоритмы:

— Метод k-ближайших соседей / k-nearest neighbors regressor (‘KNN’)

— Деревья решений / Decision Tree Regressor (‘DTR’)

— Линейный метод опорных векторов / Linear Support Vector Machine – Regression / (‘LSVR’)

— Метод опорных векторов (регрессия) / Epsilon-Support Vector Regression (‘SVR’)

Ансамблевые алгоритмы:

— AdaBoost / AdaBoost Regressor (‘ABR’) (AdaBoost = Adaptive Boosting)

— Bagging / Bagging Regressor (‘BagR’) (Bagging = Bootstrap aggregating)

— Экстра-деревья / Extra Trees Regressor (‘ETR’)

— Градиентный boosting / Gradient Boosting Regressor (‘GBR’)

— Случайный лес / Random Forest Regressor (‘RF’)

Для сравнения работы алгоритмов на наших данных была построена модель с 14 различными алгоритмами. Параметры этих алгоритмов были подобраны для наших данных, что позволит выбрать алгоритмы, подходящие для дальнейшей оптимизации. Работу алгоритма оценивали с помощью метрики R^2, полученную в результате обучения с 10-кратной перекрестной валидацией.

На рис. 3.10 представлено сравнение алгоритмов разных видов для некоторых магнитных свойств.

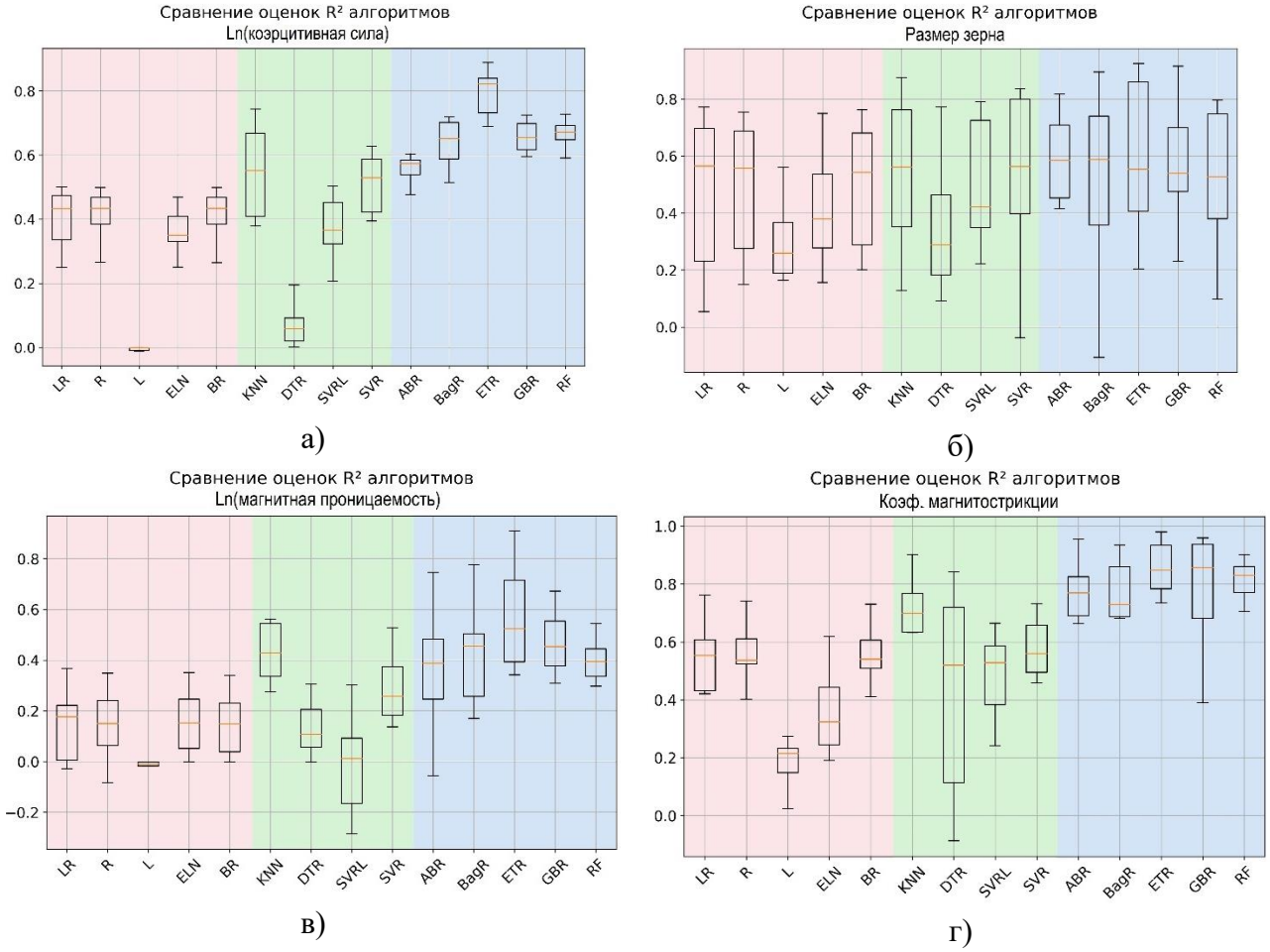


Рисунок 3.10 – Сравнение оценок R^2 различных алгоритмов для:
а – логарифма коэрцитивной силы, б – размера зерна, в – логарифма магнитной проницаемости, г – магнитострикции. Красная область – линейные алгоритмы, зеленая – нелинейные, голубая – ансамблевые

Как видно из рис. 3.10, какие-то алгоритмы описывают данные лучше, какие-то хуже.

Не менее важный параметр при выборе алгоритмов машинного обучения – гиперпараметры алгоритма и гибкость их настройки.

Проанализировав все факторы, были выбраны, следующие алгоритмы:

Метод k-ближайших соседей (KNN), Метод опорных векторов (SVR), Метод опорных векторов SVR с линейным ядром, Случайный лес (Random Forest).

3.6 Оптимизация параметров

У каждого алгоритма существуют свои параметры, изменяя которые можно добиться лучшей обобщающей способности.

Подбор оптимальных параметров происходил с помощью библиотеки Optuna [36]. Модель обучалась со значениями параметров из заданного диапазона с 10-кратной кросс-валидацией. Параметры модели с лучшей оценкой R^2 считались наилучшими и были использованы в дальнейшем.

Рассмотрим параметры каждого алгоритма, а также их влияние на результат на примере поиска оптимальных параметров для описания данных коэрцитивной силы.

Влияние на результат будем определять, фиксируя найденные на предыдущем шаге оптимальные параметры и варьируя интересующий. Влияние оцениваем с помощью метрики *MAE* (средняя абсолютная погрешность). Это оценка считается как разность предсказанного и действительного значений, деленная на количество этих значений, а затем усредненная на количество итераций в кросс валидации (на 10). Серая область на рис. 3.12 - СКО *MAE* (среднеквадратическое отклонение средней абсолютной погрешности).

к - ближайших соседей (к - Nearest Neighbors) – KNN

Алгоритм KNN – пример простого алгоритма машинного обучения с учителем для решения задач классификации и регрессии.

Суть алгоритма заключается в присваивании объекту среднего значения по k ближайшим соседям этого объекта. Пример проведения этой процедуры представлена на рис. 3.11.

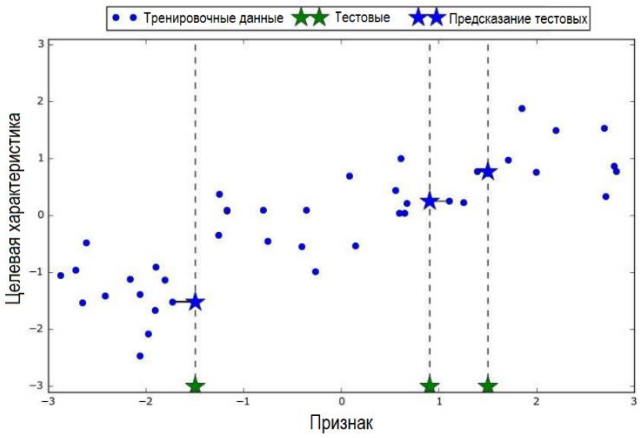


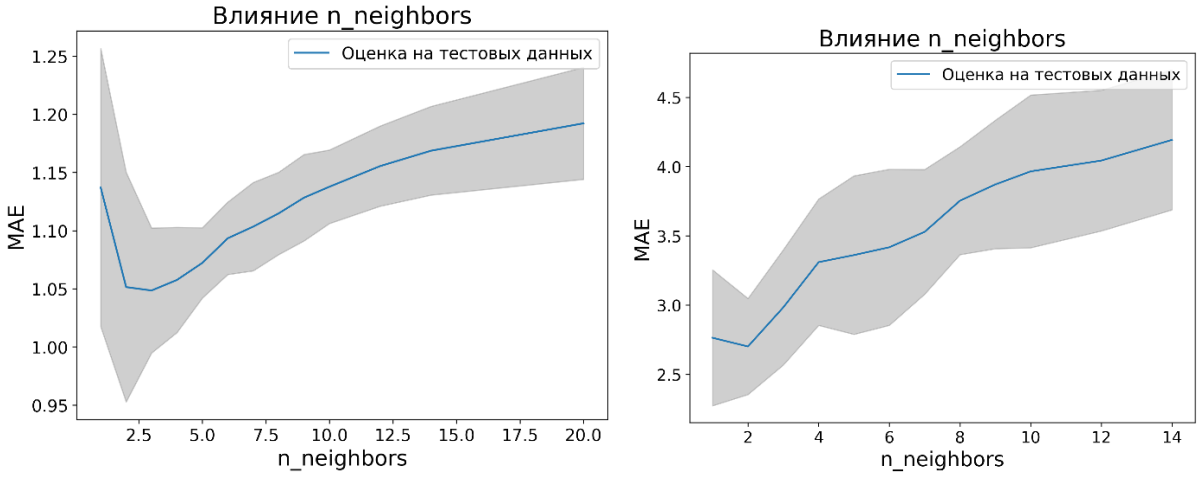
Рисунок 3.11 - Усреднение по 1 ближайшему соседу для предсказания тестовых данных [27]

Однако если соседи отдалены от объекта на разное расстояние, то у этих соседей будет различный вес для усреднения. Чем больше расстояние от объекта, тем меньший вес будет у значения соседа.

Параметры алгоритма [38]:

n_neighbors – количество ближайших соседей для усреднения. Это самый важный параметр в этом алгоритме, от него в большей степени зависит качество предсказания модели. На рис. 3.12, а видно, что существует минимум оценки *MAE* при количестве соседей *n_neighbors* = 3. А при оценке количества соседей для коэффициента магнитострикции минимум наблюдался при *n_neighbors* = 2 рис. 3.12, б.

Weights – веса соседей. Вес соседних объектов может быть одинаков, независимо от расстояния до них, а может зависеть от этого расстояния. Оценив результаты по всем целевым характеристикам, зачастую оптимальным является вес, зависящий от расстояния.



а) б)

Рисунок 3.12 - Влияние количества ближайших соседей на оценку средней абсолютной погрешности предсказания на тестовых данных:
а - коэрцитивной силы, б – коэффициента магнитострикции

Algorithm – алгоритм, с помощью которого определяются ближайшие соседи. Есть несколько алгоритмов для определения:

- *BallTree* – шаровое дерево. Поиск соседей происходит путем разбиения точек данных на набор пересекающихся гиперсфер рис. 3.13. В рамках этих сфер и происходит поиск ближайших соседей.

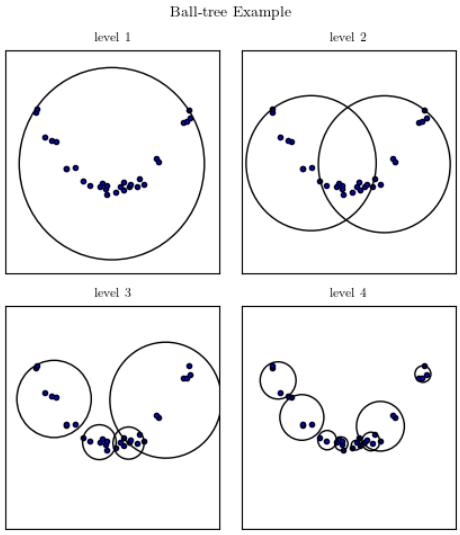


Рисунок 3.13 - гиперсферы различных уровней для построения шарового дерева

- *KDTree* - k-мерное дерево. Поиск соседей происходит путем разделения области точек данных на несколько k-мерных частей. На рис.3.14 представлен пример разбиения на 2-мерные части. Таким образом уменьшается область поиска ближайших соседей.

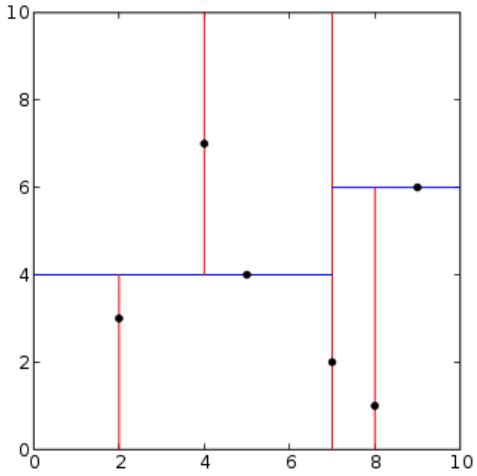


Рисунок 3.14 - Пример разделения 2-мерной области

- *Brute* – перебор всех соседей, вычисление расстояния до каждого из них, а затем поиск k ближайших.

Параметр *Algorithm* зачастую не влияет на результат определения ближайших соседей, однако очень важен для контроля времени вычислений при построении модели.

Поскольку масштабы от одного объекта до другого играет важную роль в данном алгоритме для корректной работы (чтобы вес каждого признака был равнозначен) на этапе подготовки данных мы применили масштабирование.

В результате поиска гиперпараметров были определены параметры, которые давали наибольший вклад в изменение оценки обучения, эти параметры представлены на рис. 3.15. Количество соседей для усреднения и вид веса соседей – самые важные параметры для моделей всех целевых характеристик.

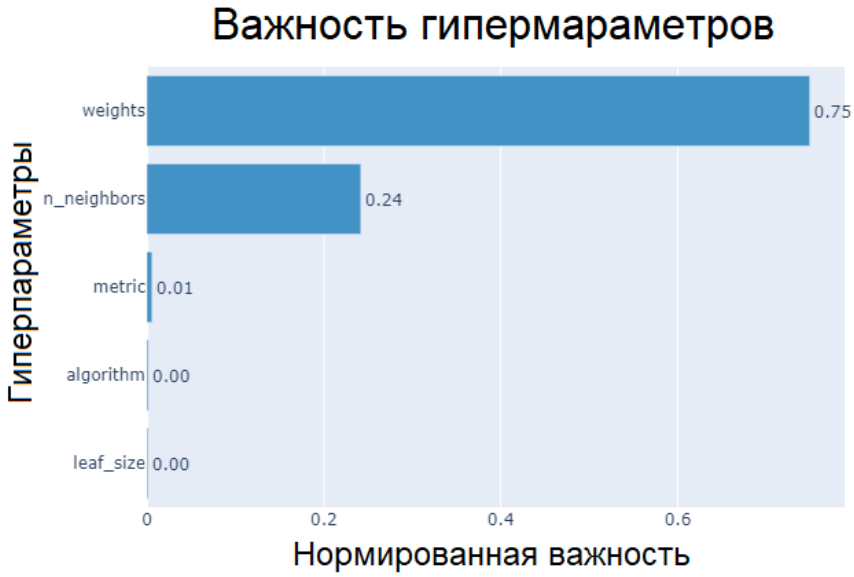


Рисунок 3.15 – Важность гиперпараметров алгоритма KNN при его оценке на тестовых данных

Метод опорных векторов (Support Vector Regression) - SVR

Суть алгоритма опорных векторов состоит в следующем:

Алгоритм определяет точки на графике как опорные вектора, строит гиперплоскость таким образом, чтобы минимизировать потери векторов, не вошедших в область нечувствительности рис. 3.16. Красная линия на рис. 3.16 – оптимальная гиперплоскость (в данном случае прямая).

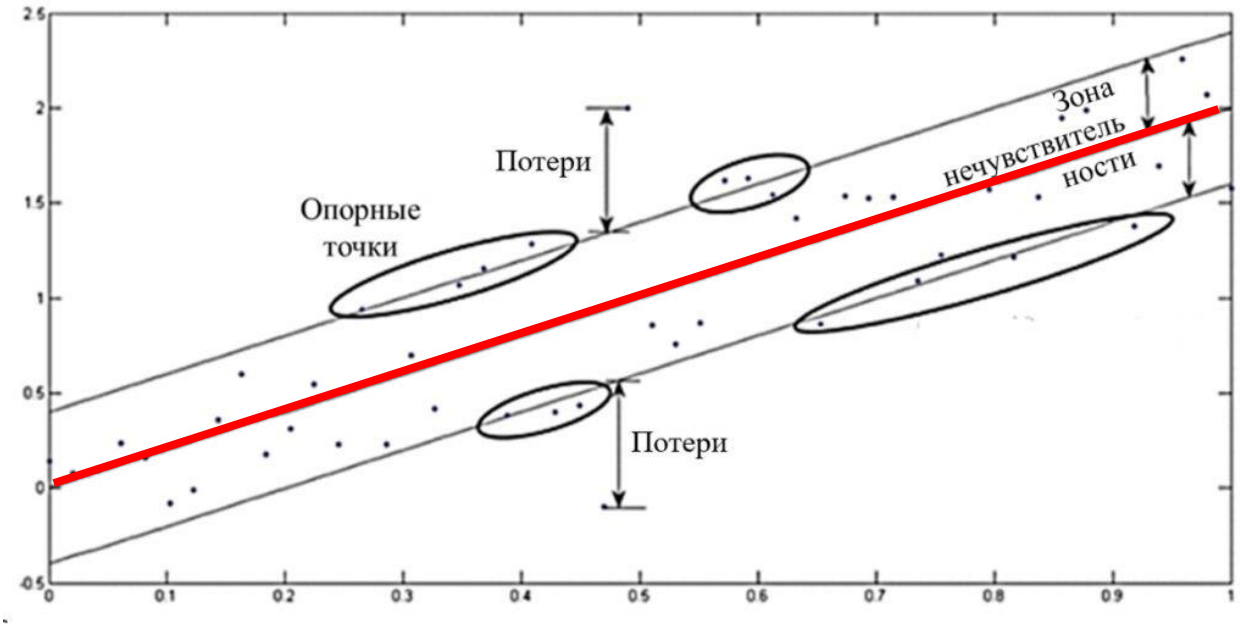
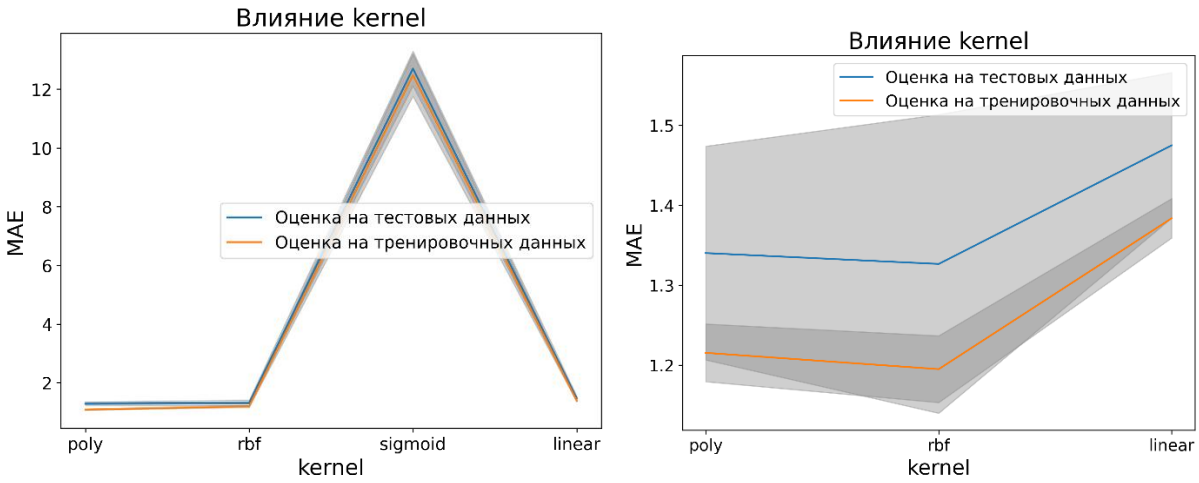


Рисунок 3.16 - Визуализация метода опорных векторов

Параметры алгоритма [39]:

kernel - Тип ядра. Функция, с помощью которой будут описываться данные. Ядро может быть линейным, полиномиальным, сигмовидным, радиально - базисной функцией (*rbf*).



а) б)

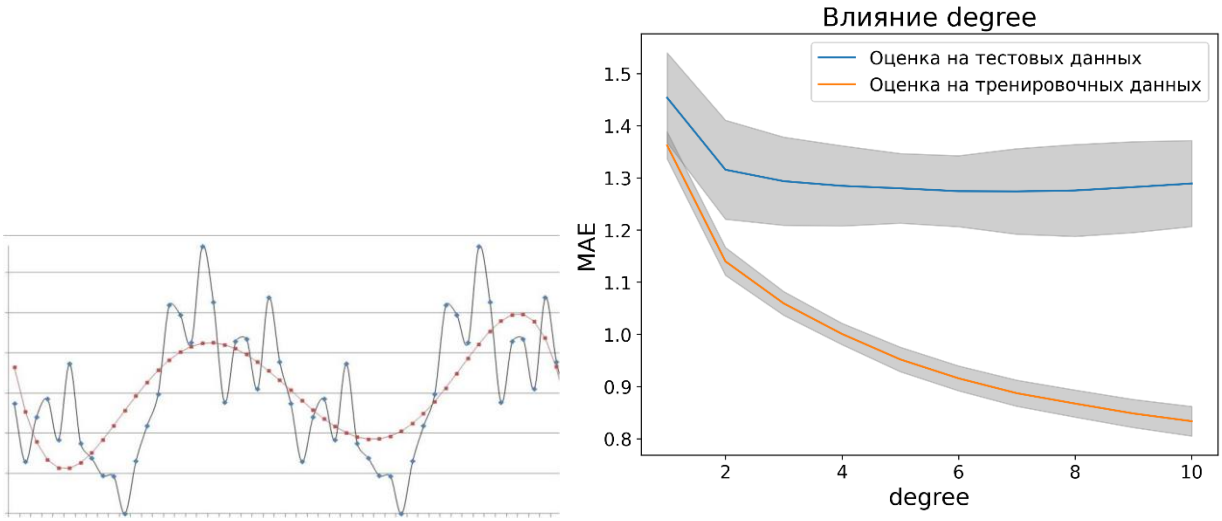
Рисунок 3.17 - Разные типы ядер алгоритма SVR:
а – все доступные ядра, б – ‘*sigmoid*’ исключен

Тип ядра существенно влияет на точность прогноза. Так, для коэрцитивной силы тип ядра ‘*sigmoid*’ показывает наихудший результат рис. 3.17, а. Для наглядности исключим

‘*sigmoid*’ из рассмотрения и рассмотрим результаты на рис. 3.17, б. Для описания данных коэрцитивной силы наиболее подходящей является полиномиальная функция за счет меньшего разброса оценки, а также ее близкое значение к минимуму.

degree - Степень полинома (только для ядра типа “*poly*”). От степени полинома зависит обучение модели. В случае если степень полинома превосходит необходимую, происходит переобучение модели, как на рис. 3.18, а. Модель точно описывает тренировочные данные, но оценка предсказания тестовых значительно больше.

По рис. 3.18, б можно определить, что оптимальной степенью является 2, поскольку оценка на тестовых данных приобретает минимальное значение.



а) б)

Рисунок 3.18 - Влияние степени полинома на результат:
а – переобучение за счет неподходящей степени полинома, б – изменение полинома в модели для коэрцитивной силы

epsilon – величина зоны нечувствительности рис. 3.16. Чем больше величина зоны нечувствительности, тем больше будет абсолютная погрешность первого значения, который окажется вне зоны чувствительности рис. 3.19. Значение между 0 и 1 в минимуме значения *MAE* на тестовых данных рис. 3.19 – является оптимальным.

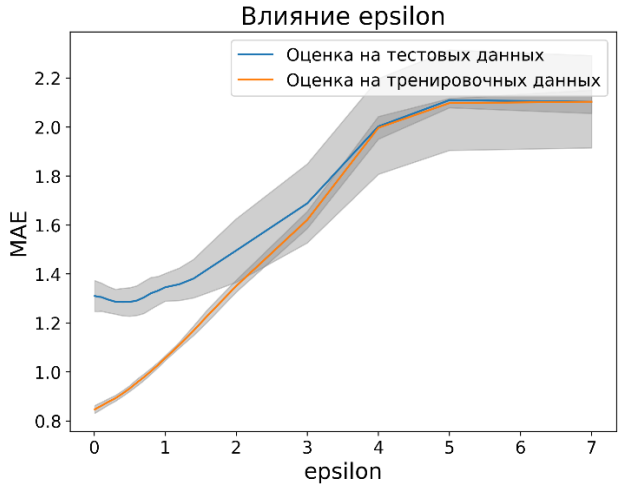


Рисунок 3.19 - Влияние параметра *epsilon*

C - Параметр регуляризации. Этот параметр помогает отрегулировать точность описания объектов обучающей выборки. Чем больше *C*, тем более витиеватой будет функция и будет лучше описывать обучающий набор данных рис. 3.20. Но при этом, будет хуже предсказывать значения тестовых данных.

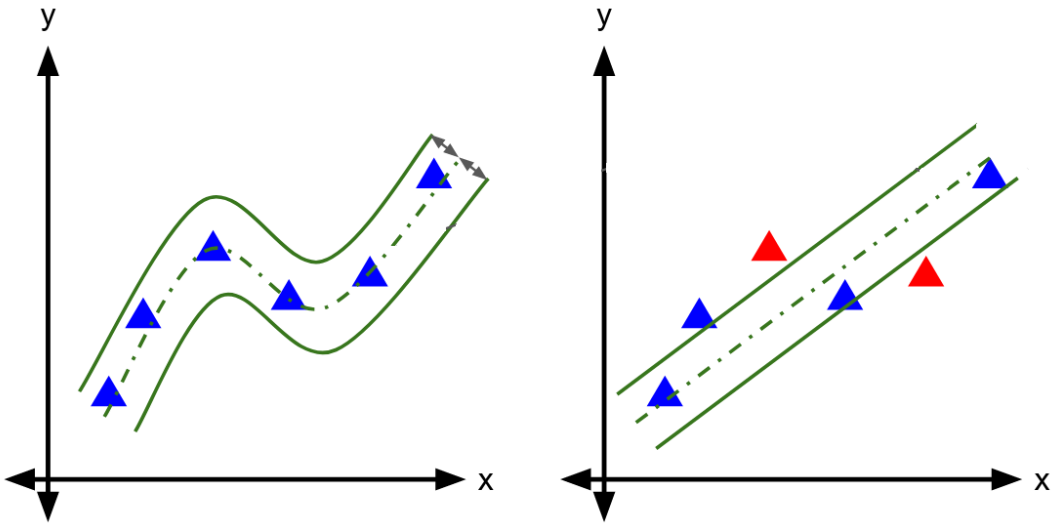


Рисунок 3.20 - Пример разных значений параметра *C*

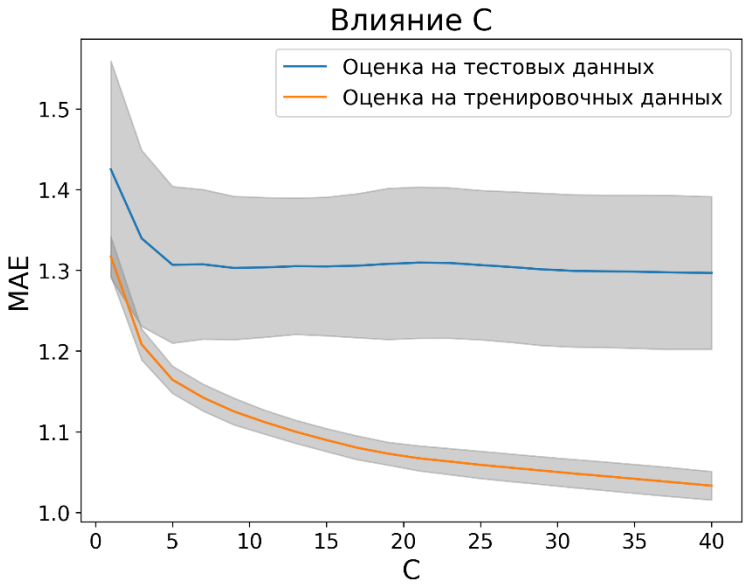


Рисунок 3.21 - Влияние параметра *C*

Как показывает рис. 3.21 модель с высоким параметром *C* склона к переобучению. Значение параметра, при котором *MAE* на тестовых данных перестает уменьшаться, является оптимальным для построения модели.

На рис. 3.22 показана важность параметров для алгоритма SVR при обучении на данных индукции насыщения. Изменение этих параметров приводило к значительному изменению оценки при обучении.

Рисунок 3.22 - Важность гиперпараметров алгоритма SVR при оценке алгоритма на обучающих данных

Случайный лес (Random Forest Regressor) – RFR

Этот алгоритм — это множество решающих деревьев рис. 3.23. В задаче регрессии ответы деревьев усредняются, таким образом можно избежать переобучение только на одном дереве.

Дерево решений строится по принципу «веток» и «листьев». Ветка дерева (признак) делиться на интервалы, сопоставляя значение характеристики с указанным неравенством, формируется многолистное дерево, где в листках находится значение целевой характеристики. Чтобы сделать прогноз необходимо спуститься по дереву до конечного листа, значение в нем и будет прогнозным значением.

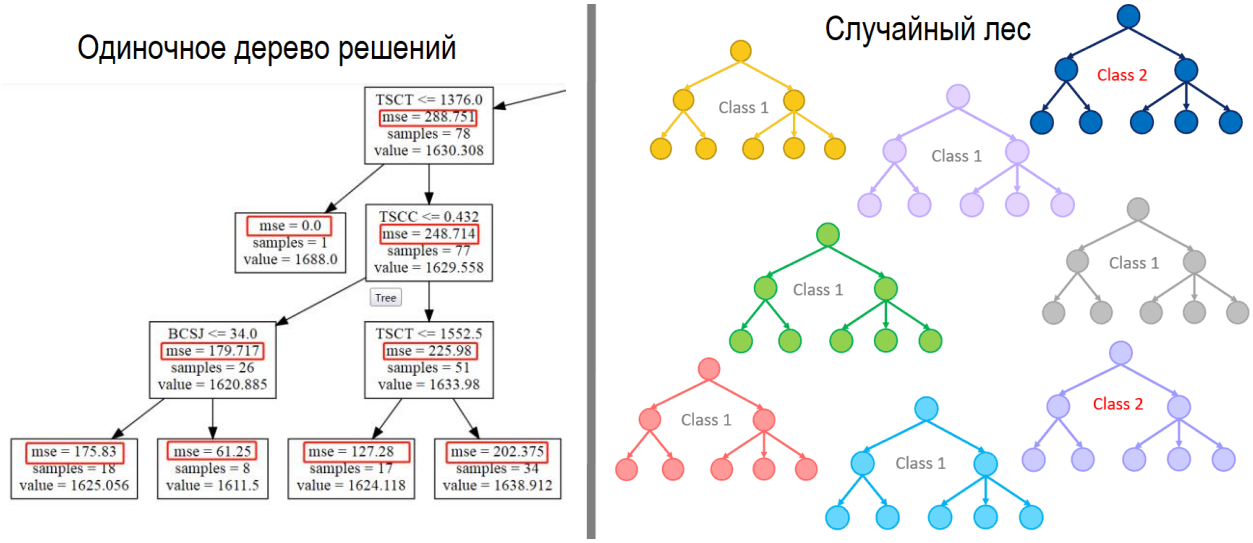


Рисунок 3.23 - Концепция алгоритма случайного леса

Параметры алгоритма [40]:

n_estimators — число «деревьев» в «случайном лесу». Чем больше деревьев, тем лучше качество модели, но, с другой стороны, тем она сложнее и требует больших мощностей для построения. В результате анализа рис. 3.24, а можно заметить, что после значения *n_estimators* = 200 изменений в графике оценок не происходит. Поэтому для оптимизации работы можно ограничиться 200 «деревьями».

max_depth — максимальная глубина дерева. Чем глубже дерево, тем лучше модель. Однако рис. 3.24, б при достижении предела глубины дерева для описания данных, не происходит дальнейшего уменьшения погрешности предсказания. Это значит, что все данные распределены по дереву с минимально доступным разделением по листьям.

min_samples_split — минимальное число объектов, необходимое для того, чтобы узел дерева мог расщепиться. Этот параметр тесно связан с максимальной глубиной дерева.

При слишком высоком значении этого параметра глубина дерева не может увеличиться из-за невозможности дальнейшего деления. На рис. 3.24, в видно, что в модели для коэрцитивной силы достаточно 2 объектов для расщепления узла.

min_samples_leaf — минимальное число объектов в листьях. Чем меньше объектов в листьях, тем точнее, но сложнее модель. Рис. 3.24, г выбираем минимальное число 2 для увеличения точности модели.

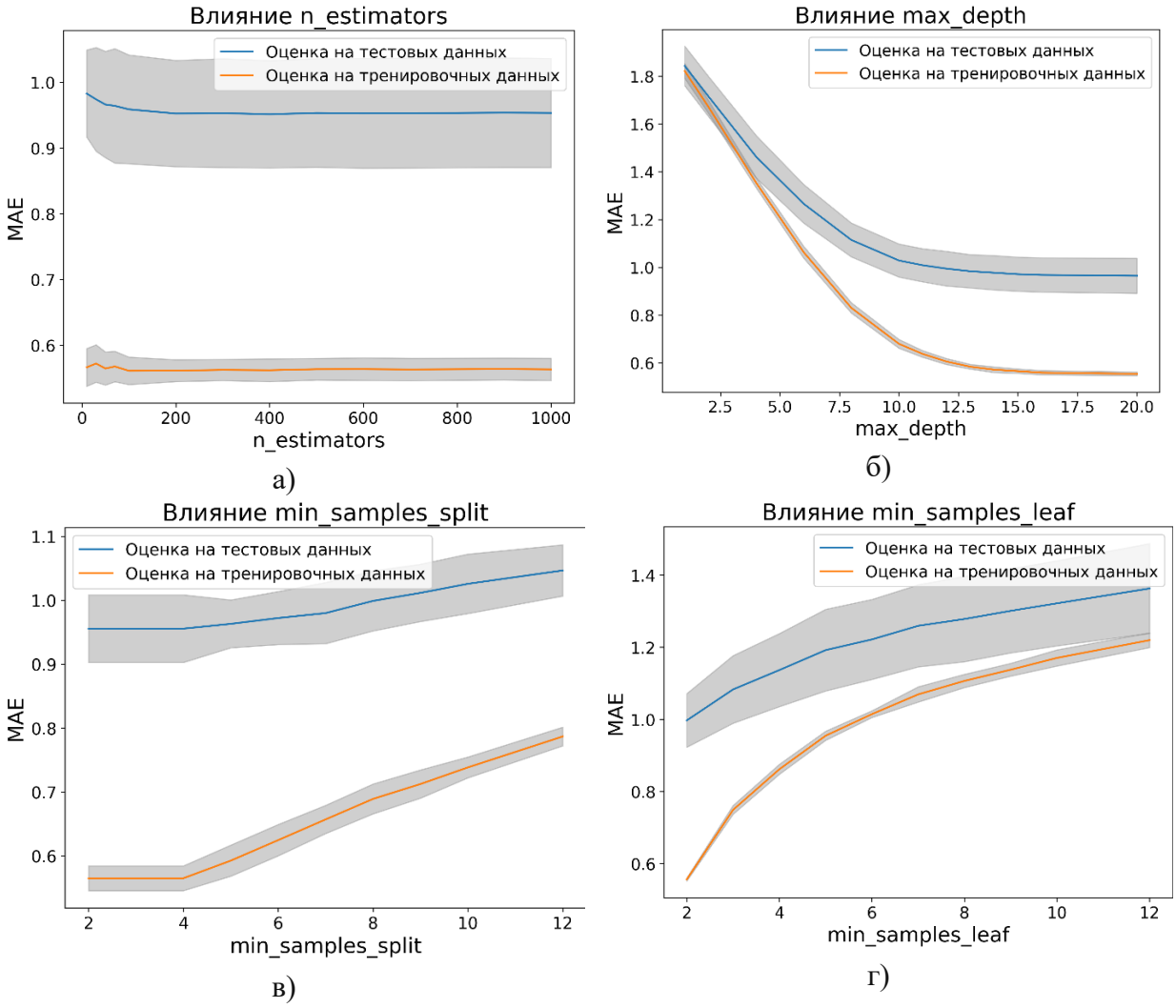


Рисунок 3.24 - Влияние параметров алгоритма на среднюю абсолютную погрешность предсказанных значений коэрцитивной силы:
а – число деревьев, б – максимальная глубина дерева, в – минимальное число объектов для разделения узла, г – минимальное число объектов в листьях.

В результате поиска оптимальных параметров были выявлены самые важные из них, они показаны на рис. 3.25. Изменение этих параметров приводило к значительному изменению оценки при обучении.

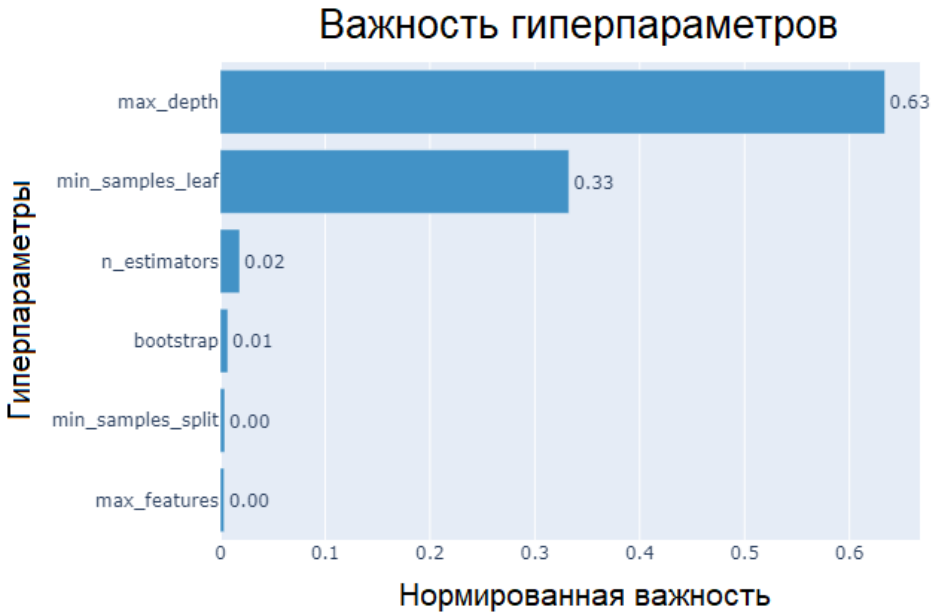


Рисунок 3.25 - Важность гиперпараметров алгоритма Random Forest при оценке алгоритма на обучающих данных

В результате поиска оптимальных параметров были определены параметры для дальнейшей работы алгоритмов.

4 Результаты и их обсуждение

После подбора оптимальных параметров алгоритмов для каждой целевой характеристики (коэффициента магнитострикции, магнитной индукции насыщения, размера зерна, коэрцитивной силы, температуры Кюри, магнитной проницаемости, электрического сопротивления) модель машинного обучения была обучена.

Оценим результат обучения по тестовым данным, которые не были доступны модели во время обучения, таким образом, мы можем оценить, насколько обученная модель может предсказывать значение целевой характеристики.

а) б)

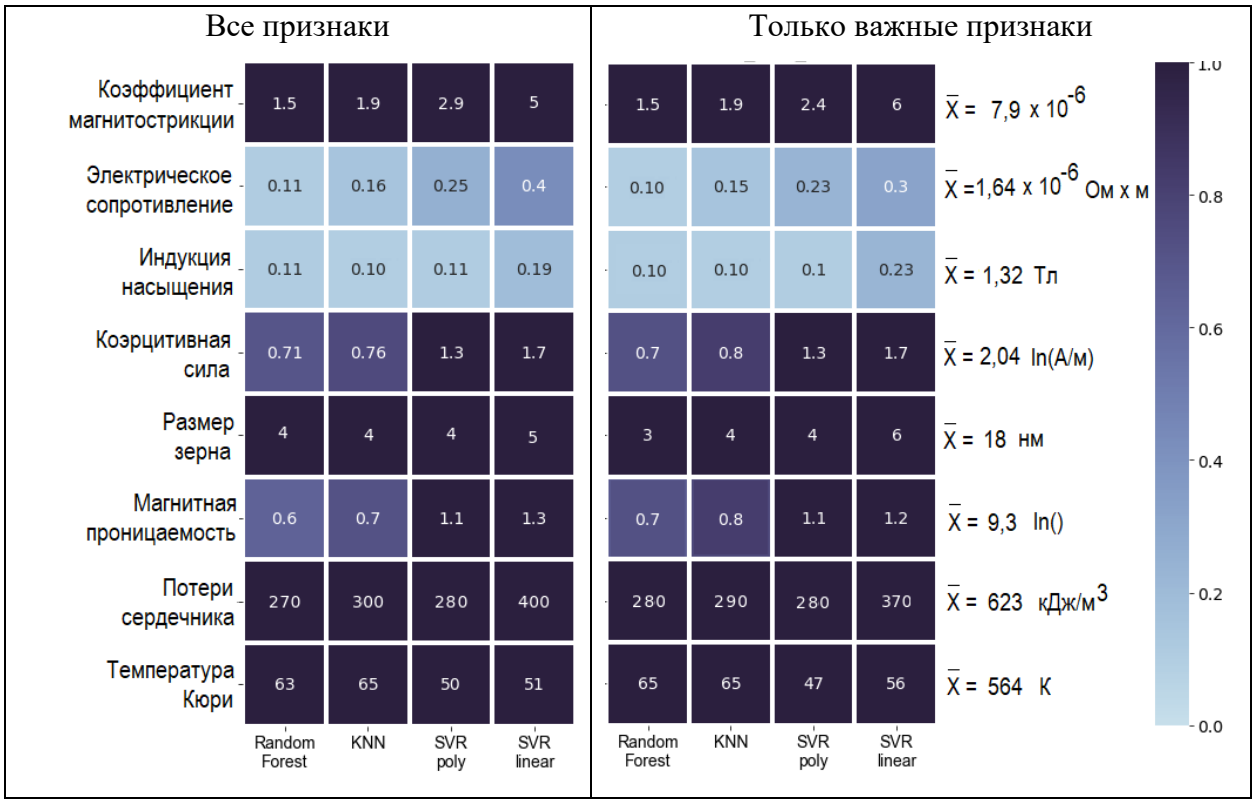
Рисунок 4.1 – оценка R^2, полученная на тестовых данных с использованием при обучении: а – всех признаков, б – только важных признаков

На рис. 4.1 представлены оценки R^2 работы моделей по прогнозу целевых характеристик. Несмотря на малое количество данных, некоторые модели показывают отличные результаты по предсказанию тестовых данных. Существует утверждение, что начиная со значения оценки R^2 ~0,8 можно считать модель машинного обучения высокого качества. С учетом этого утверждения можно обозначить 4 целевые характеристики, для которых модель имеет высокое качество, при использовании алгоритма Random Forest.

Наглядное представление о качестве предсказаний может служить оценка *MAE* на рис. 4.2. Оценка *MAE* составляет от 7 % среднего значения тестовых данных.

Кроме того, сравнивая оценки обучения с использованием всех признаков и только важных, можно отметить общее падение оценок при использовании только важных признаков, что может быть обусловлено уменьшением полезной информации для предсказания. В нашем случае, когда общее количество признаков не превышает 35, удаление незначимых признаков является нецелесообразным. Таким образом, дальнейший анализ стоит продолжать для результатов, полученных с использованием всех признаков.

Оценки R^2 для прогноза потерь сердечника, температуры Кюри, магнитной проницаемости имеют значение ниже 0,65. Такое низкое качество моделей можно объяснить малым количеством данных (для обучения и для теста) или в случае магнитной проницаемости отсутствие закономерностей с признаками, которое было найдено на этапе анализа данных.



а) б)

Рисунок 4.2 – Оценка *MAE*, полученная на тестовых данных с использованием при обучении:

а – всех признаков, б – только важных

Рассмотрим, насколько сильно искажаются графики плотности распределения предсказанных значений. На рис. 4.3 представлен график плотности распределения исходных данных.

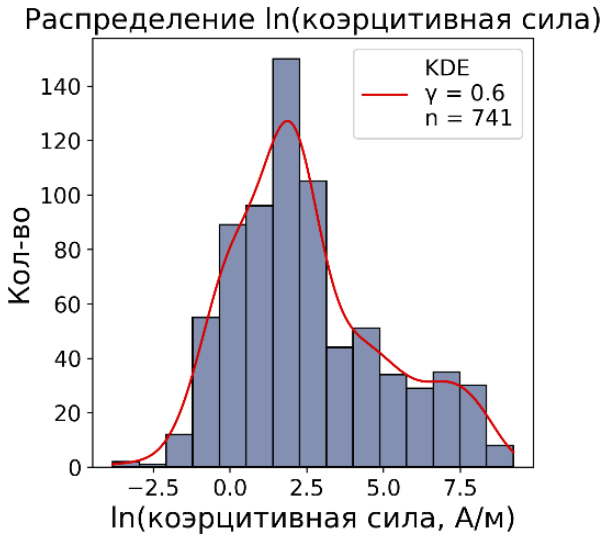
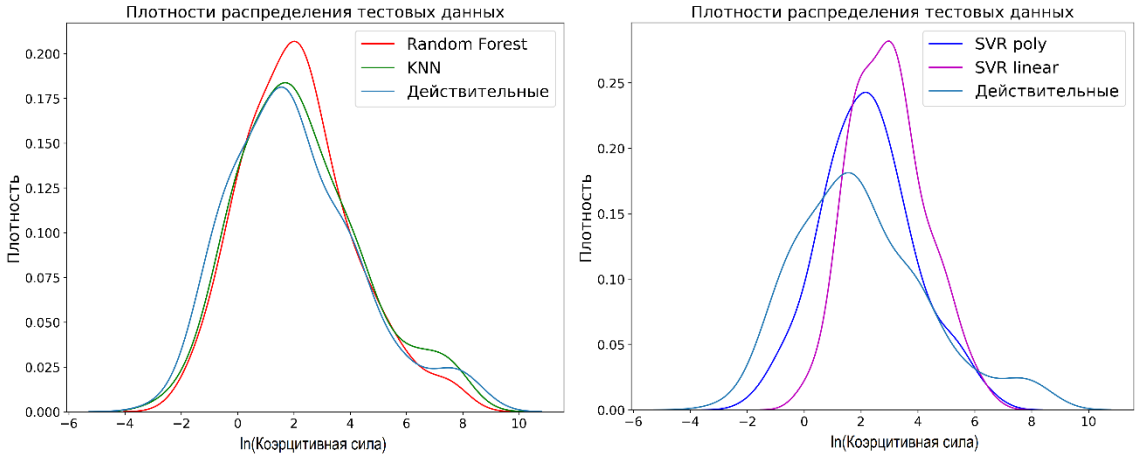


Рисунок 4.3 - Плотность распределения всех данных коэрцитивной силы.



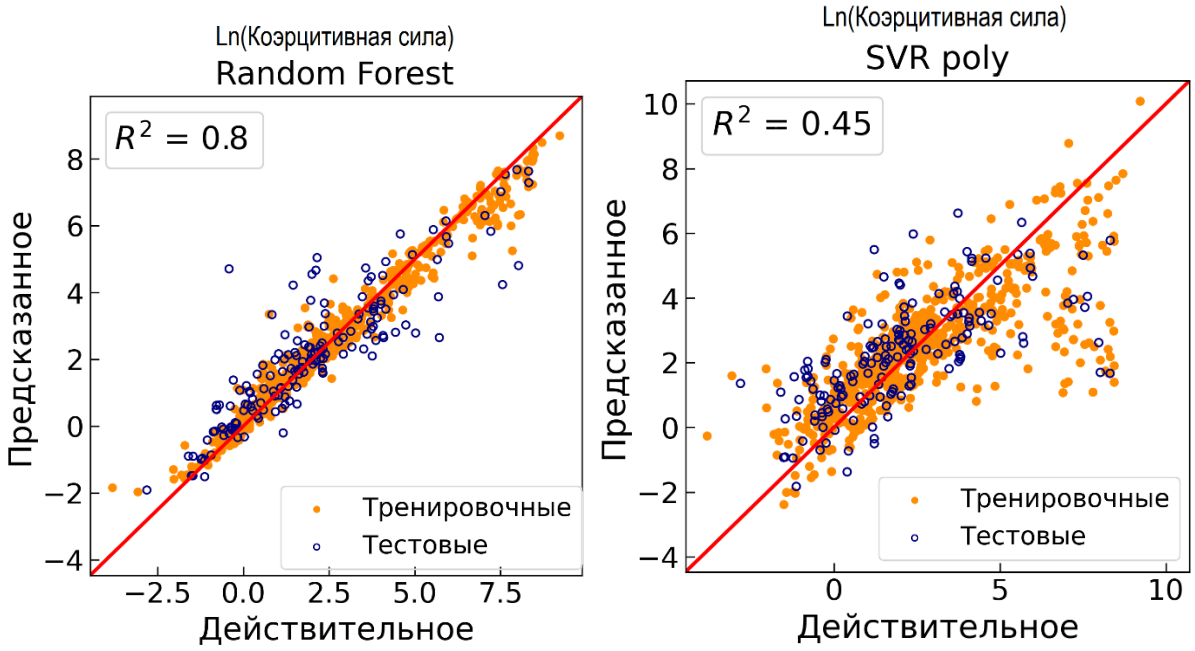
а) б)

Рисунок 4.4 – Плотность распределения действительных значений коэрцитивной силы и предсказанных алгоритмами:

а – Random Forest и KNN, б – SVR с полиномиальным и линейным ядром

На рис. 4.4 изображены графики плотности распределения для действительных и предсказанных значений коэрцитивной силы. Алгоритмы Random Forest и KNN рис. 4.4, а сохраняют исходное распределение данных, а пик плотности распределения смещается незначительно, что нельзя сказать о SVR алгоритмах (рис. 4.4, б): плотность распределения схожа с нормальным, пики плотности SVR с полиномиальным и линейными ядрами смещен в большие значения, относительно действительных.

Рассмотрим также рис. 4.5, на котором представлено соотношение действительных и предсказанных значений коэрцитивной силы. Чем уже разброс точек (чем они ближе к красной линии), тем меньше погрешность предсказания.



а) б)

Рисунок 4.5 - Соотношение действительных и предсказанных значений логарифма коэрцитивной силы:

а – алгоритма Random Forest, б – алгоритма SVR с полиномиальным ядром

Сравнение разброса точек на рис. 4.5, а и б дает представление о том, какой алгоритм лучшим образом подходит для описания данных. Делая вывод об описании коэрцитивной силы, можно отметить, как точно модель Random Forest предсказывает значения меньше 3. Существенный разброс точек наблюдается при более высоких значениях от 5 и выше. Причиной такого поведения может стать малое количество данных в этих областях.

Проанализировав рис. 4.7, а и б, где представлено то же соотношение для индукции насыщения, можно подтвердить причину существенных разбросов. Можно заметить, как для значений индукции насыщения меньше 1,25 Тл количество точек (и тренировочных, и тестовых) уменьшается в соответствии рис. 4.6, где показано распределение значений индукции насыщения.

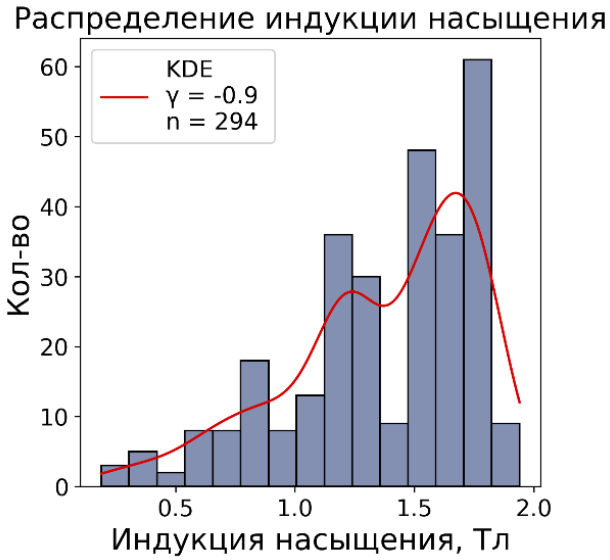
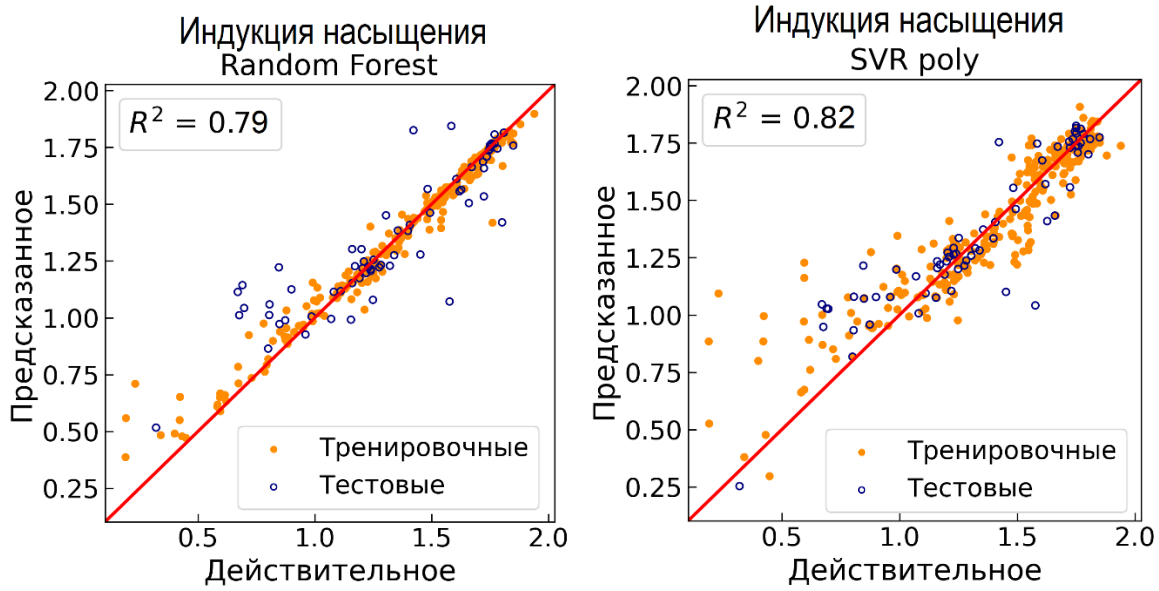


Рисунок 4.6 - Распределение значений исходных данных индукции насыщения



а) б)

Рисунок 4.7 - Соотношение действительных и предсказанных значений индукции насыщения:

а – алгоритма Random Forest, б – алгоритма SVR с полиномиальным ядром.

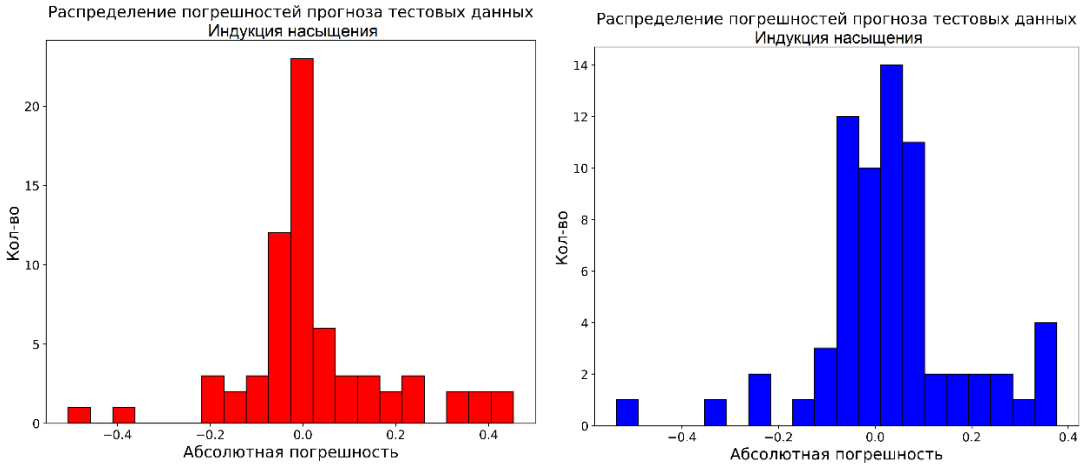
Проанализируем примеры предсказаний значений индукции насыщения в областях малых и больших значений, представленных в таблице 4.1.

Таблица 4.1 - Некоторые значения индукции насыщения и их предсказания

B_s действ. , Тл	RandomForest	Δ, Тл	KNN	Δ, Тл	SVR poly	Δ, Тл	SVR linear	Δ, Тл
0.32	0.52	0.20	0.29	-0.03	0.25	-0.07	-0.2	-0.6
0.69	1.1	0.5	0.84	0.15	1.0	0.3	1.1	0.4
0.8	0.86	0.06	0.815	0.015	0.818	0.018	0.52	-0.28
0.96	0.93	-0.03	0.85	-0.12	1.08	0.12	1.22	0.26
1.110	1.117	0.007	1.17	0.06	1.095	-0.016	1.31	0.20
1.23	1.225	-0.005	1.246	0.016	1.29	0.06	1.236	0.006
1.399	1.383	-0.016	1.3980	-0.0010	1.33	-0.06	1.36	-0.03
1.606	1.611	0.005	1.635	0.028	1.67	0.07	1.44	-0.16
1.769	1.766	-0.003	1.760	-0.009	1.7506	-0.018	1.84	0.08
1.85	1.76	-0.09	1.79	-0.06	1.77	-0.08	1.75	-0.10

Алгоритм Random Forest рис. 4.8, а и таблица 4.1 показывает отличную способность к предсказанию магнитной индукции, значение которой лежит в диапазоне 1,1 – 2,0 Тл, что дает надежду на получение прогнозов состава и условий обработки будущих нанокристаллических сплавов с индукцией, превышающей 2,0 Тл.

Распределения абсолютных погрешностей предсказаний значений индукции насыщения с помощью алгоритмов Random Forest и SVR с полиномиальным ядром представлены на рис. 4.8.



а) б)

Рисунок 4.8 - Распределение погрешностей прогноза тестовых данных с помощью алгоритмов:

а - Random Forest, б - SVR с полиномиальным ядром.

Отчетливо видна разница этих распределений. Максимум на рис. 4.8, а приходится на значения абсолютной погрешности от -0,05 Тл до +0,05 Тл. На рис. 4.8, б видно, что максимум становится шире и располагается в диапазоне от -0,1 Тл до +0,1 Тл. С помощью распределения погрешностей можно сделать вывод о вероятности точного предсказания магнитной характеристики.

Верификация полученных результатов

Произведем верификацию модели машинного обучения на актуальных данных из статей с 2010 по 2022 год выхода, информация из которых не представлена в базе данных Список этих статей представлен в приложении А. Таким образом, мы можем оценить, качество предсказания независимых валидационных данных. Такая проверка показывает, насколько полученная модель способна предсказывать магнитные характеристики новых материалов.

Актуальные данные для верификации не выбирались специально для получения высоких оценок предсказания. Некоторые целевые характеристики находились за пределами обучающих данных. Так, например, валидационные значения коэффициента магнитострикции рис. 4.9 сильно сдвинуты в отрицательную область и только частично соприкасаются с тренировочными данными. Таким образом распределены и данные потерь сердечника и электрического сопротивления. Следовательно, стоит ожидать низкие оценки качества для предсказания этих целевых характеристик.

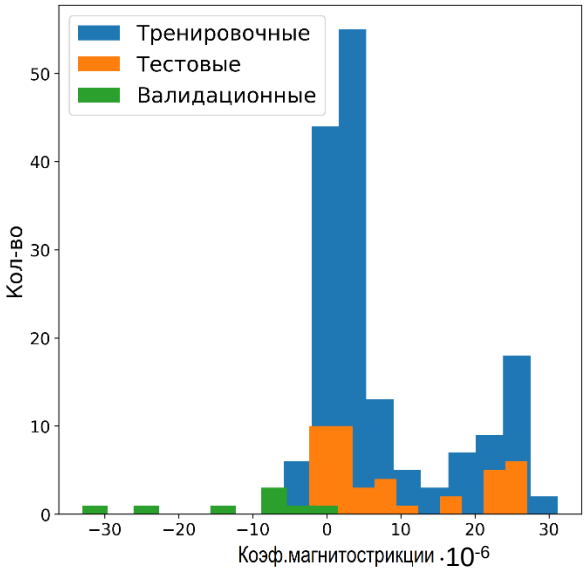


Рисунок 4.9 – Распределения коэффициента магнитострикции в тренировочной, тестовой и валидационной выборках

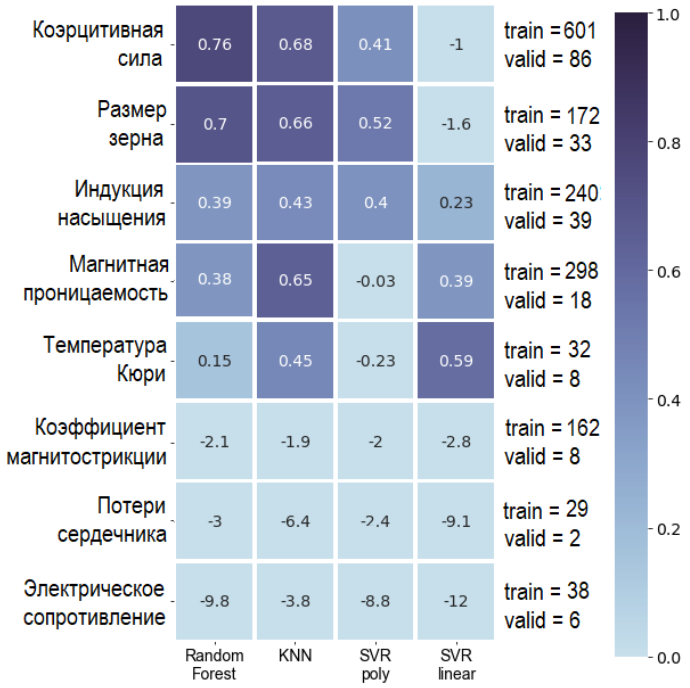
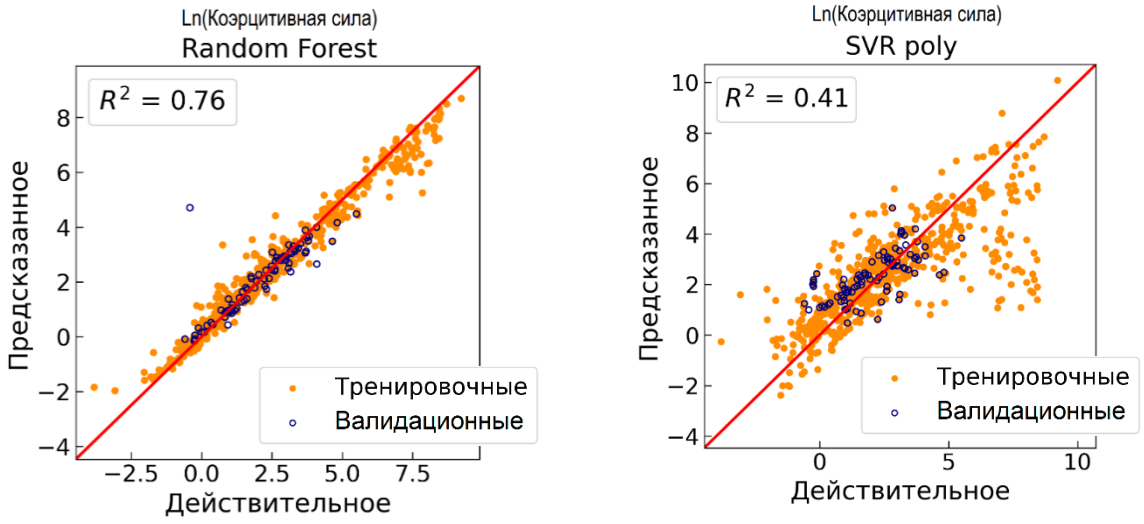


Рисунок 4.10 - Оценка R^2, полученная на валидационных данных с использованием при обучении всех признаков

Рис. 4.10 подтверждает высокое качество модели, полученное на этапе валидации модели на тестовых данных. Значения оценок R^2 ниже, чем при первичной валидации на тестовых, что связано с малым количеством данных и их различие с тренировочными данными. Однако, определение коэрцитивной силы показывает значение R^2 = 0,76 (для алгоритма Random Forest), что обусловлено большим объемом данных: 86 образцов.



а) б)

Рисунок 4.11 - Соотношение действительных и предсказанных значений логарифма коэрцитивной силы, полученных при верификации модели:

а – алгоритма Random Forest, б – алгоритма SVR с полиномиальным ядром.

На рис. 4.11 представлено соотношение действительных и предсказанных значений логарифма коэрцитивной силы, на котором можно отметить, что валидационные данные находятся в области значений, где обобщающая способность модели наилучшая. Это еще одна причина, по которой оценка R^2 на валидационных данных коэрцитивной силы остается высокой.

Результатом верификации является заключение о корректной работе модели: правильности подобранных гиперпараметров, отсутствие переобучения, а также соответствие данных требованиям алгоритмов.

Для наглядности работы модели в качестве примера возьмем образец ГМ414. Данные об образце были получены из справочных данных [6] и в процессе измерения магнитных характеристик на измерительно – вычислительном комплексе ММКС – 05. Признаки, целевые характеристики и их прогноз этого сплава представлены в таблице 4.2 и таблице 4.3.

Таблица 4.2 – Признаки образца ГМ414

Состав	Температура отжига, К	Время отжига, с	Прод./Попер. магнитное поле при отжиге	Толщина ленты, нм
$Fe_{73.5}Si_{13.5}B_9Cu_1Nb_3$	875	3600	нет/нет	25

Таблица 4.3 – Измеренные значения и прогноз магнитных свойств образца ГМ414

Целевые характеристики	Измеренные	Random Forest	KNN	SVR	SVR linear
Коэрцитивная сила, А/м	0,66	6	123	2.8	1.8
Температура Кюри, К	873	667	654	611	663
Электрическое сопротивление, $\cdot 10^{-6}$ ·Ом·м	1,25	1.73	1.62	-24	1.5
Магнитная проницаемость	34200	3467	15301	30163	42419
Коэффициент магнитострикции $\cdot 10^{-6}$	1,5	3.8	2.4	3.2	3.9
Индукция насыщения, Тл	1,225	1.16	1.09	1.27	1.2
Потери сердечника, кДж\м^3 (0,2 Тл, 100 кГц)	95	347	428	332	-239

Химический состав ГМ414 – классический состав FINEMET (см. рис. 3.6 а, б, в, г, д максимумы). По местам значений целевых характеристик в данных из базы, представленным на рис. 4.12, можно объяснить неточности в расчете магнитных свойств. Так, например, расчет коэрцитивной силы приводит к значению, превышающее

измеренное. Это происходит в результате того, что для модели значения логарифма коэрцитивной силы от 0 до 2.5 является наиболее вероятным. Поэтому при прогнозе значения логарифма коэрцитивной силы -0,4 (0,66 А/м) модель выдает значения логарифма коэрцитивной силы от 0,6 до 2 (от 1,8 до 6 А/м), руководствуясь более высокой вероятностью подобных значений характеристики для данного состава. Подобное объяснение относится и к магнитной проницаемости и электрическому сопротивлению.

В случае индукции насыщения значение для образца ГМ414 находится в области от 1,0 до 1,5 Тл, где присутствует большое количество данных, благодаря этому прогноз значения происходит с разницей ~ 5 % от измеренного значения.

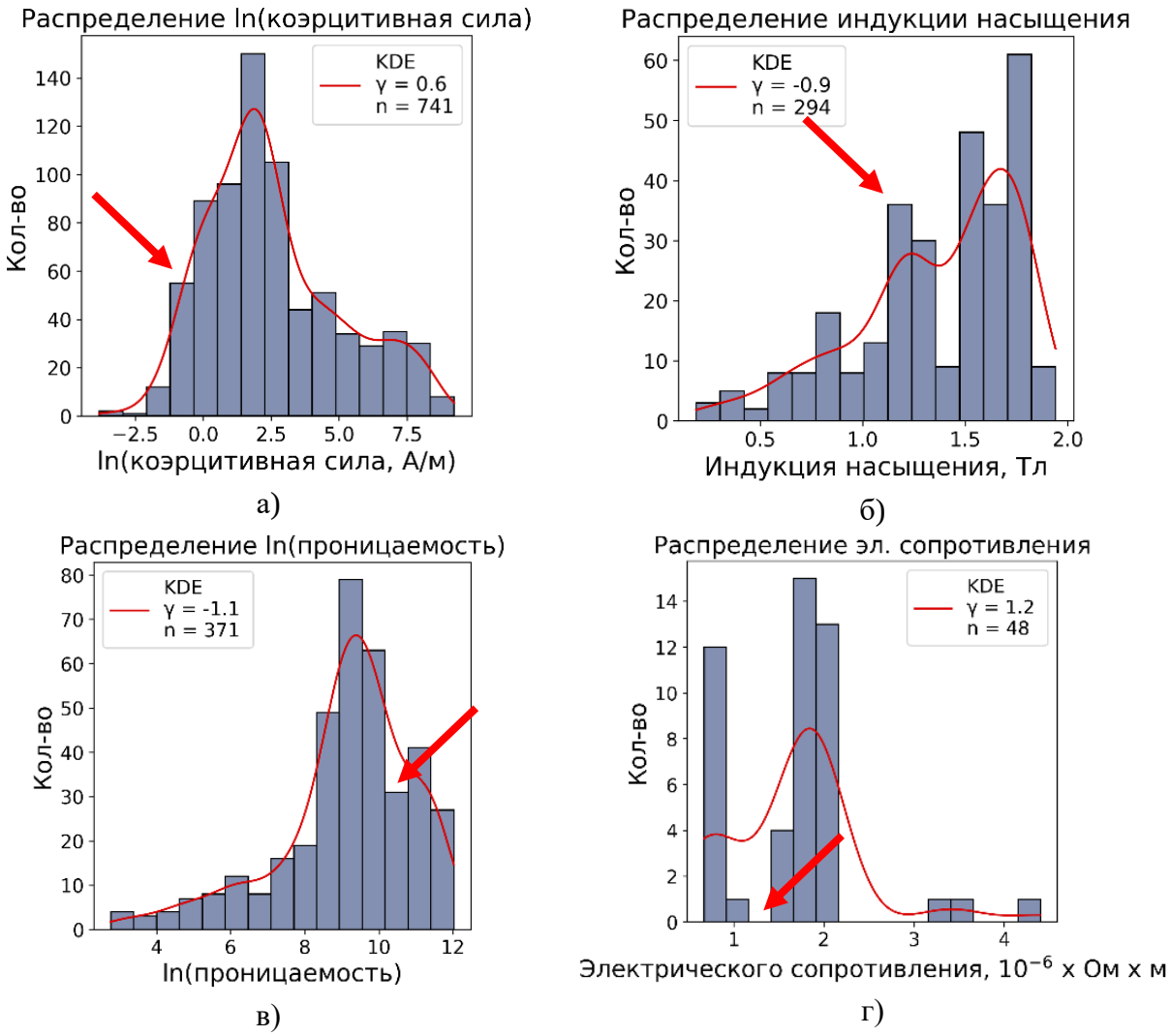


Рисунок 4.12 – Значения целевых характеристик образца ГМ414 (красные стрелочки) в распределениях данных из базы:

А – логарифм коэрцитивной силы, б – индукция насыщения, в – логарифм магнитной проницаемости, г – электрическое сопротивление.

ЗАКЛЮЧЕНИЕ

Подводя итог проделанной работы в рамках магистерской диссертации, можно выделить создание модели машинного обучения на языке программирования Python для прогноза магнитных свойств (коэрцитивной силы, коэффициента магнитострикции, индукции насыщения, размера зерна, температуры Кюри, магнитной проницаемости, электрического сопротивления, потерь на перемагничивание) по экспериментальным данным химического состава и условий обработки. Наилучший результат предсказания получается при использовании алгоритма Random Forest, так коэффициенты детерминации R^2 для коэффициента магнитострикции, индукции насыщения, размера зерна, коэрцитивной силы, электрического сопротивления, магнитной проницаемости имеют значения выше 0,65.

В ходе работы удалось не только создать модель машинного обучения для, но и провести валидацию этой модели:

Валидация экспериментальных данных из литературы, поступающих на вход. Процесс анализа заключался не только в формальной проверке наличия данных и правильности их записи, также были оценены пропуски данных, соответствие данных цели исследования, распределения признаков и целевых характеристик, а также были изучены корреляции признаков и целевых характеристик.

Валидация выбранных алгоритмов и их параметров с целью соблюдения требования о корректной обобщающей способности модели (отсутствие недо- и переобучения).

Первичная валидация модели на тестовых данных. По результатам проверки можно сделать вывод об отсутствии переобучения модели, поскольку она способна описывать незнакомые данные с оценкой R^2 0,93 (в случае коэффициента магнитострикции). Также было выявлено, что алгоритм Random Forest по оценкам R^2 имеет лучшее качество предсказаний магнитных характеристик.

Верификация полученных результатов на актуальных данных из статей, опубликованных с 2010 по 2022 год. В результате верификации удалось установить важность объема и разнообразия данных при обучении для прогноза магнитных характеристик.

Подводя итог, важно отметить полезность объединения исследователей, публикующих статьи, для занесения в стандартизированные базы данных информации об исследованных образцах. Это существенно облегчает сбор информации для развития машинного обучения и исследований, использующих методы машинного обучения. Создание стандартизированных баз данных, а также совершенствование и создание новых алгоритмов работы с данными и машинного обучения – важная задача метрологии в «Big Data».

БЛАГОДАРНОСТИ

Выражаю свою благодарность кандидату физико-математических наук, доценту кафедры магнетизма и магнитных наноматериалов Катаеву Василию Анатольевичу за руководство в выполнении магистерской диссертации, создание оптимальных условий для самореализации, развития профессиональных навыков и приобретения знаний в области физики магнитных явлений.

Также благодарю кандидата физико-математических наук, доцента кафедры магнетизма и магнитных наноматериалов Болячкина Антона Сергеевича за руководство в выполнении магистерской диссертации, за переданные знания и консультирование в области машинного обучения.

Отдельной благодарности заслуживают авторы статьи [2] Wang Y., Tian Y., Kirk T., Laris O., Ross Jr. J.H., Noebe R.D.,Keylin V., Arroyave R. за предоставленную базу данных с помощью которой проводилось данное исследование.

Также благодарю коллег к.ф.-м.н. Степанову Елену Александровну, к.ф.-м.н. Волегова Алексея Сергеевича, к.ф.-м.н. Незнахина Дмитрия Сергеевича и д.ф.-м.н. Васьковского Владимира Олеговича за ценные замечания и проявленный интерес к теме работы.

СПИСОК ЛИТЕРАТУРЫ

1 Jha R., Chakraborti N., Diercks D. R., Stebner A. P., Ciobanu C. V. Combined machine learning and CALPHAD approach for discovering processing-structure relationships in soft magnetic alloys // Computational Materials Science. – 2018. – V. 150, – P. 202–211.

2 Wang Y., Tian Y., Kirk T., Laris O., Ross Jr. J.H., Noebe R.D.,Keylin V., Arroyave R. Accelerated design of Fe-based soft magnetic materials using machine learning and stochastic optimization // Acta Materialia. – 2020. – V. 194. – P. 144–155.

3 Li X., Shan G., Shek C. H. Machine learning prediction of magnetic properties of Fe-based metallic glasses considering glass forming ability // Journal of Materials Science and Technology. – 2022. – V. 103. – P. 113–120.

4 Lu Z., Chen X., Liu X., Lin D., Wu Y., Zhang Y., Wang H., Jiang S., Li H., Wang X., Lu, Z. Interpretable machine-learning strategy for soft-magnetic property and thermal stability in Fe-based metallic glasses // Npj Computational Materials. – 2020. – V. 6, №. 187. – P. 1–9.

5 Debnath B., Vinoth A., Mukherjee M., & Datta S. Designing Fe-based high entropy alloy-a machine learning approach // Paper presented at the IOP Conference Series: Materials Science and Engineering. – 2020. – V. 912, №. 5. – P. 1–7.

6 Стародубцев Ю. Н. Нанокристаллические магнитомягкие материалы // Компоненты и технологии, 2007. № 4, с. 240–242.

7 Zhun Li, Kefu Yao, Deren Li, Xiaojun Ni, Zhichao Lu. Core loss analysis of Finemet type nanocrystalline alloy ribbon with different thickness // Progress in Natural Science: Materials International. – 2017. – V. 27, №. 5. – P. 588–592.

8 Tsepelev V. S., Starodubtsev Y. N. Nanocrystalline Soft Magnetic Iron-Based Materials from Liquid State to Ready Product // Nanomaterials. – 2021. – V. 11, №.1. – P. 1-39.

9 Petrescu L.G, Petrescu M.C, Cazacu E, Constantinescu C.D. Estimation of Energy Losses in Nanocrystalline FINEMET Alloys Working at High Frequency // Materials. – 2021. – V. 14, №. 24. – P. 1–13.

10 Ding J, Xu Hongjie, Shi Zhiguang, Li Xuan, Zhang Tao. Effect of primary α-Fe on soft magnetic properties of FeCuNbSiB amorphous/nanocrystalline alloy // Journal of Non-Crystalline Solids. – 2021. – V. 571. – P. 1–6.

11 Sinha A.K., Singh M.N., Upadhyay A., Satalkar M., Shah M., Ghodke N., Kane S.N., Varga L.K. A correlation between the magnetic and structural properties of isochronally annealed Cu-free FINEMET alloy with composition $Fe_{72}B_{19.2}Si_{4.8}Nb_4$ // Appl. Phys. A. – 2015. – V. 118, №. 1. – P. 291–299.

12 Mushnikov N.V., Potapov A.P., Shishkin D.A. Magnetic properties and structure of nanocrystalline FINEMET alloys with various iron contents // Phys. Metals Metallogr. – 2015. – V. 116. – P. 663–670.

13 Xie Z.Y., Wang Z., Han Y. Effect of Ge addition on structure and soft magnetic properties of Si-rich Fe-based nanocrystalline alloys // J. Alloys Compd. – 2016. – V. 672. P. 332–335.

14 Kwapulinski P., Rasek J., Stokosa Z., Haneczok G. Optimisation of soft magnetic properties in Fe-Cu-X-$Si_{13}B_9$ (X= Cr, Mo, Zr) amorphous alloys // J. Magn. Magn. Mater. – 2001. – V. 234, № 2. – P. 218–226.

15 Franco V., Conde C.F., Conde A. Magnetic properties and nanocrystallization of a $Fe_{63.5}Cr_{10}Si_{13.5}B_9Cu_1Nb_3$ alloy // J. Magn. Magn. Mater. – 1999. – V. 203. – P. 60–62.

16 Lim S.H., Pi W.K., Noh T.H., Kim H.J., Kang I.K. Effects of Al on the magnetic properties of nanocrystalline $Fe_{73.5}Cu_1Nb_3Si_{13.5}B_9$ alloys // J. Appl. Phys. – 1993. – V. 73, № 10. – P. 6591–6593.

17 Yuchen M., Zhenghou Z., Hui Z. Microstructures and soft magnetic properties of $Fe_{73.5-x}Cu_1Nb_3Si_{13.5}B_9Y_x$ (x = 0–1.5) alloys // Results in Physics. – 2022. – V. 34. – P. 1–6.

18 Meng Xiao, Zhigang Zheng, Li Ji, Xin Liu, Zhaoguo Qiu, Dechang Zeng, The role of V and Mo on crystallization process and magnetic properties of FeSiBCuNb alloys using in wide frequency scale // Journal of Non-Crystalline Solids. – 2019. – V. 521. – P. 1–6.

19 Yoshizawa Y. Magnetic properties of Fe-Cu-M-Si-B (M = Cr, V, Mo, Nb, Ta, W) alloys // Materials Science and Engineering. – 1991. – V. A133. – P. 176–179.

20 Vlasak G., Kaczkowski Z., Svec P., Duhaj P. Influence of heat treatment on magnetostrictions of Finemet $Fe_{73.5}\,Cu_1\,Nb_3\,Si_{13.5}\,B_9$ // Materials Science and Engineering: A. – 1997. – V. 226. – P. 749–752.

21 Gu L. Y., Wang S. Z., Bai X. F., Zhang X., Kong Q. K., Li X. S., & Xue Z. Y. Study on properties and heat treatment process of $Fe_{75.9}Cu_1Si_{13}B_8Nb_{1.5}Mo_{0.5}Dy_{0.1}$ // Nano. – 2021. – V. 16, №11.

22 Han J., Kwon S., Sohn S., Schroers J., & Choi-Yim H. Optimum soft magnetic properties of the Fe–Si–B–Nb–Cu alloy achieved by heat treatment and tailoring b/si ratio // Metals. – 2020. – V. 10, № 10. – P. 1–8.

23 Ningning Shen, Zhengxu Dou, Yuluo Li, Kuang Lv, Yidong Wu, Fushan Li, Xidong Hui. Effect of Fe content on crystallization behavior and soft magnetic properties in FINEMET-type alloys // Materials Letters. – 2021. – V. 305. – P. 1–4.

24 Kwon S., Kim S., & Yim H. Improvement of saturation magnetic flux density in Fe–Si–B–Nb–Cu nanocomposite alloys by magnetic field annealing // Current Applied Physics. – 2020. – V. 20, № 1. – P. 37–42.

25 Mikhalitsyna E., Zakharchuk I., Soboleva E., Geydt P., Kataev V., Lepalovskij V., & Lähderanta E. Heat treatment effect on magnetic microstructure of $Fe_{73.9}Cu_1Nb_3Si_{13.2}B_{8.9}$ thin films // EPJ Web of Conferences, 2018. – V. 185.

26 Михалицына Е. А. Магнитная анизотропия и гистерезисные свойства аморфных и нанокристаллических пленок Fe-M-Cu-Si-B (M: Nb, NbMo, W) : дис… канд. физ.-мат. наук. – Екатеринбург, 2018. – 165 с.

27 A. Muller, S. Guido. Introducing to Machine Learning with Python. – Sebastopol, O'Reilly Media, Inc, 2016 – 338 p.

28 Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining // Journal of Data Warehousing. – 2000. – V. 5. – P. 13–22.

29 ГОСТ Р 58776 – 2019 Средства мониторинга поведения и прогнозирования намерений людей. Термины и определения. – введ. 2020–09–01. – М. : Стандартинформ, 2020. – 6 с.

30 ГОСТ Р 58777 – 2019 Воздушный транспорт. Аэропорты. Технические средства досмотра. Методика определения показателей качества распознавания незаконных вложений по теневым рентгеновским изображениям. – введ. 2020–09–01. – М. : Стандартинформ, 2020. – 12 с.

31 ГОСТ Р 59879 – 2021 Эргономика. Проектирование и применение испытаний речевых технологий. Методика определения показателей качества распознавания голосовых команд управления. – введ. 2022–03–01. – М. : Российский институт стандартизации, 2021. – 20 с.

32 ГОСТ 33707 – 2016 (ISO/IEC 2382:2015) Информационные технологии. Словарь. – введ. 2017–09–01. – М. : Стандартинформ, 2016. – 201 с.

33 Могильников И. Валидация моделей машинного обучения [Электронный ресурс] // Хабр. URL: https://habr.com/ru/post/453558/ (дата обращения 11.03.2022)

34 Морозова О.А. К вопросу определения метрик качества данных // Управленческие науки в современном мире. – 2018. – Т. 1. – № 1. – С. 180–184.

35 Перекрестная проверка: оценка производительности [Электронный ресурс] // Scikit-learn Машинное обучение в Python. URL: https://scikit-learn.ru/3-1-cross-validation-evaluating-estimator-performance/ (дата обращения 01.05.2022)

36 Takuya A., Shotaro S., Toshihiko Y., Takeru O., and Masanori K. Optuna: A Next-generation Hyperparameter Optimization Framework. // 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). – 2019. – P. 2623–2631.

37 Дьяконов А. «Анализ малых данных». Глава 8. Метрики качества [Электронный ресурс] // АНАЛИЗ МАЛЫХ ДАННЫХ. URL: https://alexanderdya-

konov.files.wordpress.com/2018/10/book_08_metrics_12_blog1.pdf. (Дата обращения: 10.03.2022)

38 KNeighborsRegressor [Электронный ресурс] // Scikit-learn Машинное обучение в Python URL : https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html (дата обращения 05.03.2022)

39 SVR [Электронный ресурс] // Scikit-learn Машинное обучение в Python URL : https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html (дата обращения 05.03.2022)

40 RandomForestFegressor [Электронный ресурс] // Scikit-learn Машинное обучение в Python URL : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (дата обращения 05.03.2022)

ПРИЛОЖЕНИЕ А

1 Xie Z., Wang Z., Han Y. Effect of Ge addition on structure and soft magnetic properties of Si-rich Fe-based nanocrystalline alloys // Journal of Alloys and Compounds. – 2016. – V. 672. – P. 332–335.

2 Xie L., Wang A., Yue S., He A., Chang C., Li Q., Wang X., Liu C. Significant improvement of soft magnetic properties for Fe-based nanocrystalline alloys by inhibiting surface crystallization via a magnetic field assisted melt-spinning process // Journal of Magnetism and Magnetic Materials. – 2019. -V. 483. – P. 158–163.

3 Zhu Q., Chen Z., Zhang S., Li Q., Jiang Y., Wu P., Zhang K. Improving soft magnetic properties in FINEMET-like alloys with Ga addition // Journal of Magnetism and Magnetic Materials. – 2019. -V. 487. – P. 1–6.

4 Liu Y., Li J., Sun Y., He A., Dong Y., Wang Y. Effect of annealing temperature on magnetic properties and corrosion resistance of $Fe_{75.8}Si_{12}B_8Nb_{2.6}Cu_{0.6}P_1$ alloy // Journal of Materials Research and Technology. – 2021. -V. 15. – P. 3880–3894.

5 Jia X., Zhang B., Zhang W., Dong Y., Li J., He A., Li R. Direct synthesis of Fe-Si-B-C-Cu nanocrystalline alloys with superior soft magnetic properties and ductile by melt-spinning // Journal of Materials Science & Technology. – 2022. -V. 108. – P. 186–195.

6 Xiao M., Zheng Z., Ji L., Liu X., Qiu Z., Zeng D. The role of V and Mo on crystallization process and magnetic properties of FeSiBCuNb alloys using in wide frequency scale // Journal of Non-Crystalline Solids. – 2019. -V. 521. – P. 1–6.

7 Ding J., Xu H., Shi Z., Li X., Zhang T. Effect of primary α-Fe on soft magnetic properties of FeCuNbSiB amorphous/nanocrystalline alloy // Journal of Non-Crystalline Solids. – 2021. -V. 571. – P. 1–6.

8 Jia Y., Wang Z., Wang F., Zhang L., Duan H. Effect of Ti on structure and soft magnetic properties of Si-rich Finemet-type nanocrystalline $Fe_{73.5}Cu_1Nb_{3-x}Si_{17.5}B_5Ti_x$ alloys // Materials Research Bulletin. – 2018. -V. 106. – P. 296–300.

9 Shen N., Dou Z., Li Y., Lv K., Wu Y., Li F., Hui X. Effect of Fe content on crystallization behavior and soft magnetic properties in FINEMET-type alloys // Materials Letters. – 2021. -V. 305. – P. 1–4.

10 Shivaee H.A., Hosseini H., Lotfabad E.M., Roostaie S. Study of nanocrystallization in FINEMET alloy by active screen plasma nitriding // Journal of Alloys and Compounds. – 2010. -V. 491. – P. 487–494.

11 Moya J.A. Nanocrystals and amorphous matrix phase studies of Finemet-like alloys containing Ge // Journal of Magnetism and Magnetic Materials. – 2010. -V. 322. – P. 1784–1792.