Международный Университет Информационных Технологий

УДК: 004.852:004.75

На правах рукописи

БАКИРОВА ГУЛЬНАЗ САЙЛАУОВНА

Разработка моделей и методов с применением федеративного машинного обучения

8D06102 - Компьютерная и программная инженерия

Диссертация на соискание степени
доктора философии (PhD)

Научный консультант:
кандидат технических наук,
профессор,
Бектемысова Г. У.
Зарубежный консультант
PhD, ассоциированный профессор
Нор'ашикин Бинти Али

Республика Казахстан
Алматы, 2026

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ

В данной диссертации использованы ссылки на следующие стандарты:

Инструкция по подготовке диссертации и реферата, Высшая аттестационная комиссия
Министерства образования и науки Республики Казахстан от 28 сентября 2004 г. № 377-3

ГОСТ 7.32-2001. Отчет о научно-исследовательских работе. Структура и правила оформления.

ГОСТ 7.1-2003. Библиографическая запись. Библиографическое описание. Общие требования и правила составления.

СТ РК 34.005-2002. Информационные технологии. Основные термины и определения (первое издание).

СТ РК. 34.015-2002. Информационные технологии. Набор стандартов для автоматизированных систем. Техническое задание на создание ИС (первое издание).

СТ РК 34.027-2006. Информационные технологии. Классификация программных средств (первое издание).

СТ РК 34.014-2002. Информационные технологии. Набор стандартов для автоматизированных систем. Автоматизированные системы. Термины и определения.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ИИ	Искусственный Интеллект
МО	Машинное Обучение
ФО	Федеративное Обучение
ГВО	Горизонтальное Федеративное Обучение
ВФО	Вертикальное Федеративное Обучение
Non-IID	не независимо и не идентично распределенные данные
GDPR	General Data Protection Regulation (Общий регламент защиты данных)
RF	Random Forest
SMPC	Secure Multi-Party Computation
HIPAA	Health Insurance Portability and Accountability Act
IoT	Internet of Things
FedAvg	Federated Averaging
FedSGD	Federated Stochastic Gradient Descent
FedProx	Federated Proximal
LSTM	Long Short-Term Memory
Lasso-регрессия	Least Absolute Shrinkage and Selection Operator
MSE	Mean Squared Error
R^2	коэффициент детерминации

ОПРЕДЕЛЕНИЯ

Федеративное обучение – это подход в машинном обучении, где обучение модели выполняется на распределенных устройствах, не передавая данные клиентов на центральный сервер, где происходит агрегация локальных обновлений модели.

Горизонтальное федеративное обучение – это вид федеративного обучения, в котором у клиентов одинаковое пространство признаков, при этом разные наборы обьектов данных.

Вертикальное федеративное обучение – это вид федеративного обучения, в котором у клиентов разное пространство признаков, при этом пересекающие или одинаковые обьектов наблюдений.

Cross-Silo федеративное обучение - данный тип федеративного обучения, которая применяется между организациями (университетами, больницами) при котором каждый клиент имеет значительный обьем данных и стабильными вычислительными ресурсами.

Non-IID данные – это распределение данных между клиентами федеративного обучения, которое влияет на сходимость модели и снижает качество обобщения.

Random Forest (Случайный лес) – это ансамблевый алгоритм машинного обучения, который строит множество решающих деревьев и агрегирует их предсказаний, чтобы повысить точность и устойчивость модели.

Ансамблевые методы – это класс методов машинного обучения, где для получения более точных и устойчивых предсказаний применяют комбинацию нескольких базовых моделей.

Психоэмоциональное выгорание – это состояние эмоционального, физического и когнитивного истощения, возникающее в результате длительного воздействия стрессовых факторов, характеризующееся снижением мотивации, продуктивности и эмоционального благополучия.

Опросник психоэмоционального выгорания – это адаптированный инструмент для сбора данных, определяющий количественную оценку уровня эмоционального истощения, деперсонализации и снижения личных достижений.

One-Hot Encoding – это метод кодирования категориальных признаков, при котором каждому уникальному значению сопоставляется бинарный вектор.

Среднеквадратичная ошибка (MSE) – это метрика оценки качества регрессионной модели, вычисляемая как среднее значение квадратов разностей между фактическими и предсказанными значениями.

Коэффициент детерминации (R^2) - это метрика, которая показывает долю дисперсии целевой переменной, объясняемую моделью.

Агрегация моделей – это процесс обьединения локальных обновлений моделей клиентов федеративного обучения с целью формирования глобальной модели.

FedAvg (Federated Averaging) – это базовый алгоритм федеративного обучения, который основан на усреднении параметров локальных моделей клиентов с учетом обьема их данных.

FedSGD –это алгоритм федеративного обучения, где клиенты передают градиенты после каждой итерации локального обучения.

FedProx – это модификация алгоритма FedAvg, который вводит проксимальный член в функцию потерь, повышающая устойчивость обучения non-IID данных.

FedOpt – это класс алгоритмов федеративного обучения, где используются адаптивные оптимизаторы на стороне сервера для улучшения сходимости глобальной модели.

ВВЕДЕНИЕ

Актуальность темы исследования. Образовательная среда переходит в эпоху глобальной цифровизации, поэтому в этой области идет устойчивое развитие инструментов анализа распределенных и конфиденциальных данных, в связи с этим идет активный рост научного и практического интереса к методам федеративного машинного обучения. Ужесточение требований к защите конфиденциальных данных при сборе больших обьемов информации ограничивает возможности централизованной обработки. Сохранение конфиденциальности данных, отражающие учебную активность, поведенческие характеристики и психоэмоциональное состояние студентов, собираемых в современных образовательных системах, являются особенно важной задачей.

Потенциал федеративного обучения при работе с распределенными данными подтверждаются результатами зарубежных исследований. Точность моделей, которая сопоставляется с централизованным обучением, обеспечивается базовым механизмом федеративной агрегации. Нестабильная сходимость и снижение качества глобальной модели характеризуется в условиях несбалансированных и гетерогенных данных, что отражено в исследованиях. В научной литературе предложены усовершенствованные схемы федеративного обучения для преодоления упомянутых ограничений. Подходы, основанные на серверной оптимизации, направлены на повышение устойчивости и ускорение сходимости процесса обучения и методы, использующие проксимальную регуляризацию, позволяют ограничить расхождение локальных моделей и стабилизировать обучение в условиях гетерогенности данных. Возможность достижения высокой точности без передачи исходных данных между клиентами отражены в дополнительных исследованиях в области горизонтального федеративного обучения.

Вопросам защиты персональных данных уделяется особое внимание в научной литературе. Федеративное обучение является одним из наиболее перспективных методов для образовательных и социальных систем, при этом позволяя комбинировать аналитическую эффективность и соблюдение требований информационной безопасности.

Анализ исследований показал, что в области федеративного обучения большинство работ преимущественно ориентировано на дифференцируемые модели и нейросетевые архитектуры. Применение федеративных подходов к недифференцируемым алгоритмам машинного обучения, в частности к ансамблевым моделям типа Random Forest. Разработка методов федеративной агрегации, основанных на статистических представлениях локальных моделей, остаются недостаточно изученными. Практическая реализация для анализа психоэмоционального состояния студентов, основанных на реальных гетерогенных данных образовательных учреждений, имеет ограниченное количество упоминаний в научной литературе.

Задача разработки и практическая реализация методов федеративного машинного обучения для недифференцируемых моделей для анализа распределенных конфиденциальных данных студентов является актуальной. Решением данной задачи является обеспечение высокого уровня защиты персональной информации, повышение устойчивости и качества глобальных моделей в условиях гетерогенных данных, развитие цифровой трансформации образовательной среды и внедрение интеллектуальных аналитических систем при строгом соблюдении требований информационной безопасности.

Целью диссертационного исследования является разработка моделей и методов федеративного машинного обучения и их практическая реализация для анализа распределенных конфиденциальных данных при обеспечении требований защиты персональной информации и устойчивости обучения в условиях гетерогенных данных.

Задачи исследования:

1. Реализовать систему сбора, предварительной обработки и согласования признакового пространства распределенных данных студентов.
2. Разработать методы федеративной агрегации для недифферецируемых ансамблевых моделей на основе статистических представлений локальных моделей.
3. Разработать архитектуру федеративного машинного обучения для анализа распределенных конфиденциальных данных с локальной обработкой информации без передачи исходных данных на сервер.
4. Провести экспериментальную оценку эффективности, устойчивости и сходимости разработанных моделей в условиях гетерогенных (non-IID) данных.

Обьектом исследования является образовательная среда, в которой формируются и используются распределенные конфиденциальные данные студентов для анализа и прогнозирования их психоэмоционального состояния с применением методов федеративного машинного обучения.

Предметом исследования являются методы и алгоритмы федеративного машинного обучения, предназначенные для анализа и прогнозирования психоэмоционального состояния студентов на распределенных и гетерогенных данных при обеспечении конфиденциальности информации.

Методы исследования:

1. Методы системного анализа и проектирования при разработке архитектуры федеративной обучающей системы и веб-платформы.
2. Методы предварительной обработки и преобразования данных: очистка, валидация, обработка аномалий, кодирование признаков и преобразование временных характеристик.
3. Методы ансамблевого и федеративного машинного обучения, основанные на агрегировании статистических представлений локальных моделей.

4. Методы математической статистики и вычислительного эксперимента для оценки качества, устойчивости и сходимости глобальной модели.

Научная новизна исследования заключается в том, что:

1. Разработан и реализован подход к модификации и расширению федеративного машинного обучения для недифференцируемой ансамблевой модели Random Forest Regression на основе замены классической агрегации параметров модели на агрегирование статистических представлений локальных моделей, а именно векторов важностей признаков, что обеспечивает устойчивость процесса обучения в условиях гетерогенных non-IID распределенных данных.

2. Разработаны модели и методы федеративного машинного обучения для анализа распределенных конфиденциальных данных, обеспечивающие локальную обработку информации, устойчивость глобальной модели и сохранение конфиденциальности данных при работе в реальных условиях распределенной образовательной среды.

3. Предложена архитектура федеративного обучения, обеспечивающая формирование глобальной модели на основе локальных обновлений без передачи исходных данных клиентов на сервер, что позволяет соблюдать требования конфиденциальности и защиты персональной информации.

Теоретическая значимость исследования: изучение вопросов распределенного анализа данных на основе теории машинного обучения. Мы применили FedAvg, FedOpt, FedProx в качестве алгоритмов федеративного обучения на децентрализованных наборах данных и Random Forest Regression в качестве ансамблевой модели машинного обучения для предсказаний, а также классические методы статистического анализа, и предварительная обработка данных являются ключевыми методами, которые применялись в исследовании. Работа шла по стандартным шагам машинного обучения от подготовки данных до оценки результатов. Реализовано было на языке Python с использованием библиотек sckit-learn, pandas, NumPy matpotlib инструменты, которые упростили проведение экспериментов и сделали визуализацию наглядной.

Практическая значимость исследования: исследование направлено на развитие и расширение методов федеративного машинного обучения за счет разработки и реализации подхода, адаптированного для недифференцируемой ансамблевой модели Random Forest Regression. В работе предложена замена классической агрегации параметров модели на агрегирование статистических представлений локальных моделей в виде векторов важностей признаков, что обеспечивает устойчивость процесса обучения в условиях гетерогенных non-IID распределенных данных.

Разработана архитектура федеративного обучения, позволяющая формировать глобальную модель на основе локальных обновлений без передачи исходных данных клиентов на сервер, что обеспечивает соблюдение требований конфиденциальности и защиты персональной информации.

Основные положения, выносимые на защиту:

- Разработана и реализована архитектура федеративного машинного обучения для анализа распределенных и конфиденциальных данных студентов, обеспечивающая обучение локальных моделей без передачи исходной информации на сервер и соответствующая требованиям защиты персональных данных.
- Выполнена адаптация алгоритмов федеративного обучения FedAvg, FedOpt, FedProx к недифференцируемой модели Random Forest Regression путем замены агрегации весов на агрегацию статистических представлений – векторов важностей признаков, что позволило обеспечить устойчивость обучения при гетерогенных non-IID данных.
- Проведен сравнительный анализ эффективности алгоритмов федеративной агрегации на реальных распределенных данных, отражающих психоэмоциональное состояние студентов, показавший, что FedProx обеспечивает наибольшую стабильность при non-IID распределениях, а FedOpt ускоряет сходимость глобальной модели.

В ходе диссертационного исследования были получены следующие научные результаты, которые определяют его научную новизну:

- Реализована система сбора, предварительной обработки и согласования признакового пространства распределенных данных студентов.
- Разработаны методы федеративной агрегации для недифференцируемых ансамблевых модели на основе статистических представлений локальных моделей.
- Разработана архитектура федеративного машинного обучения для анализа распределенных конфиденциальных данных с локальной обработкой информации без передачи исходных данных на сервер.
- Проведена экспериментальная оценка эффективности, устойчивости и сходимости разработанных моделей в гетерогенных non-IID данных.

Основные результаты работы были представлены и обсуждены на семинарах кафедры «Компьютерная инженерия» АО МУИТ (2022–2026 гг.), университете Tenaga Nasional Малайзия (2024–2026 гг.). По теме диссертации опубликованы 8 публикаций: 1 статья - в изданиях, индексируемых в базе Web of Science и Scopus; 4 статьи - в журналах, рецензируемом Комитетом по обеспечению качества в сфере науки и высшего образования МНВО РК, 2 статьи - в Международном Журнале Информационных и Коммуникационных технологий, Спецвыпуск. 1 статья - в Вестнике КазНТУ, Computing & Engineering.

1. Бакирова Г. С., Бектемысова Г. У., Нор'ашикин Бинти Али. Federated machine learning for monitoring student mental health in Kazakhstan. International Journal of Advanced Computer Science and Applications. E-ISSN: 2156-5570 P-ISSN: 2158-107X. Volume 16 Issue 10. 2025. стр. 212–220. CiteScore - 2. Наивысший процентиль–47%. Квартиль - Q3. DOI: https://doi.org/doi:10.14569/IJACSA.2025.0161022

2. Бакирова Г. С., Бектемысова Г. У. Анализ алгоритмов федеративного обучения. Вестник КазАТК, Вестник Казахской Академии Транспорта и Коммуникации им. М.Тынышпаева, ISSN 2790-5802, - Том 131 №2, -2024г. С.: 297-304 DOI: https://doi.org/10.52167/1609-1817-2024-131-2-297-304.

3. Бакирова Г. С., Бектемысова Г. У. Сравнительный анализ алгоритмов объединенного машинного обучения. Scientific Journal of Astana IT University. ISSN (P): 2707-9031 ISSN (E): 2707-904X, Volume 17, March 2024, P.:57-67. DOI: https://doi.org/doi:10.37943/17BVCN7579

4. Бакирова Г. С., Бектемысова Г. У., Ермуханбетова Ш., Шынторе Г. А., Умуткулов Д. Б., Мангышева Ж. С. Анализ актуальности и перспективы применения федеративного обучения. Вестник НИА РК, №2(92), 2024, С.: 56-65 DOI: https://doi.org/10.52167/1609-1817-2024-131-2-297-304

5. Бектемысова Г. У., Ахмер Е. Ж., Сабденов А., Бакирова Г. С., Разработка модели для классификации документа (на примере паспортов). Вестник КазАТК, Вестник Казахской Академии Транспорта и Коммуникации им. М.Тынышпаева, ISSN 2790-5802, - Том 136 №1, -2025г. -С. 393-401 DOI: https://doi.org/10.52167/1609-1817-2025-136-1-393-401

6. Бакирова Г. С., Бектемысова Г. У. Обзор о распределенном и федеративном машинном обучении для моделей больших данных. Международный Журнал Информационных и Коммуникационных технологий, Спецвыпуск 2023. ISSN 2708-2032 (print) ISSN 2708-2040 (online), Спецвыпуск 2023. стр. 89-96. https://ydf.iitu.edu.kz/files/26-37-PB.pdf

7. Бакирова Г.С., Бектемысова Г.У. Analysis of probable threats in the use of federated learning and their protection methods . Международный Журнал Информационных и Коммуникационных технологий, Спецвыпуск 2024. ISSN 2708-2032 (print) ISSN 2708-2040 (online), стр. 64-69. https://ydf.iitu.edu.kz/files/%D0%A1%D0%BF%D0%B5%D1%86%D0%B2%D1%8B%D0%BF%D1%83%D1%81%D0%BAYDF2024.pdf

8. Бакирова Г.С., Бектемысова Г.У, Шайкемелов Г. Applications in federated machine learning. Вестник КазНТУ, Computing & Engineering, Том 1 №3, - 30.09. 2023г. С.: 25-28, DOI: https://doi.org/10.51301/ce.2023.i3.05

9. Бакирова Г. С., Бектемысова Г. У. Свидетельство на право охраны программы для ЭВМ № 62530 Республики Казахстан. Система анализа для выявления признаков психоэмоционального выгорания студентов с использованием методов федеративного обучения. заявка 26.09.2025; публикация 30.09.2025.

Результаты. В процессе работы над исследованием изучены актуальные подходы федеративного машинного обучения, уделено особое внимание алгоритмам FedAvg, FedOpt, FedProx, учитывая вопросы устойчивости и сходимости моделей при гетерогенных non-IID распределенных данных. Анализ научных публикаций показал, что федеративное обучение является перспективным инструментом для обработки конфиденциальной информации, однако его применение к недифференцируемым моделям остается недостаточно изученным.

Архитектура федеративного обучения для анализа распределенных данных, которая отражает психоэмоциональное состояние студентов, была разработана и реализована в рамках исследования. Рассмотрим более подробно архитектуру, которая состоит из локальных моделей, обучающиеся на стороне клиентов, с применением Random Forest Regression, которая обеспечивает высокую точность прогнозирования и устойчивость к шуму данных и глобальных моделей, обучающиеся на стороне сервера, что означает серверная часть выполняет агрегацию без доступа к исходным данным.

Сбор данных осуществлялся через веб-платформу, которую отправляли всем клиентам, в течение 2 месяцев с 1 сентября по 30 октября. Информационная база включала данные о питании, физической активности, сна и отдыха, психоэмоционального состояния, а также временные и идентификационные атрибуты. Данные собирались от нескольких учебных заведений, что позволило смоделировать условия реальной гетерогенности. Предварительная подготовка данных включала очистку некорректных записей, фильтрацию выбросов, преобразование временных признаков в числовой формат и кодирование категориального идентификатора student_id методом One-Hot Encoding. Мы провели проверку согласованности признакового пространства между локальными наборами данных, которые необходимы для корректной федеративной агрегации.

Алгоритмы федеративного обучения FedAvg, FedOpt, FedProx были адаптированы к недифференцируемой модели Random Forest Regression путем замены агрегации весов модели на агрегацию статистических представлений, т.е. векторов важностей признаков, что позволяет сохранить математическую логику федеративного обучения и обеспечивает устойчивость глобальной модели.

Сравнительный анализ алгоритмов федеративной агрегации провели по регрессионным и классификационным метрикам. В результате показал, что обеспечивает базовый уровень качества и будет использоваться как эталон, при этом его сходимость ухудшается при выраженной гетерогенности данных, в свою очередь демонстрирует ускоренную сходимость за счет применения серверной оптимизации, но обеспечил наиболее стабильное обучение и минимальные колебания качества. Динамика точности и функции потерь глобальных моделей была проанализирована по раундам федеративного обучения. Все алгоритмы показывают устойчивый рост точности и монотонное снижение Centralized Loss, что демонстрирует корректную сходимость и отсутствие переобучения. Следовательно, цели и задачи диссертационного исследования были достигнуты и полученные результаты подтверждают эффективность применения адаптированных методов федеративного обучения для анализа психоэмоционального состояния студентов на распределенных конфиденциальных данных и обладают как научной, так и практической значимостью.

В первой главе рассмотрены основные подходы к федеративному обучению и обоснован выбор синхронного горизонтального Cross-Silo

федеративного обучения. Использование единого пространство признаков позволило корректно агрегировать локальные результаты без передачи исходных данных. Архитектура обеспечила устойчивую сходимость глобальной модели, сохранение конфиденциальности и эффективность обучения в условиях гетерогенных данных, что подтверждает применимость предложенного решения в образовательной среде.

Во второй главе рассмотрены алгоритмы и архитектуры федеративного обучения, показано, что базовые методы FedAvg и FedSGD, обеспечивают конфиденциальность данных, но теряют устойчивость при non-IID распределениях. В свою очередь FedOpt FedProx повышают стабильность и ускоряют сходимость за счет проксимальной регуляризации и серверной оптимизации. В результате выбор FedAvg, FedOpt и FedProx является наиболее сбалансированным решением для анализа распределенных конфиденциальных данных.

Третья глава посвящена описанию веб-платформы сбора данных о питании, физической активности, сне и психоэмоциональном состоянии студентов, а также использованию опросника Maslach Burnout Inventory (MBI) и его адаптации для исследования. Рассмотрены методы обработки non-IID данных, обеспечения конфиденциальности и математические модели анализа психоэмоционального состояния. Описаны алгоритмы прогнозирования выгорания, принципы кодирования данных и автоматизированного анализа факторов, влияющих на уровень выгорания. Также обозначены ограничения исследования и перспективы дальнейшего развития предложенного подхода.

В четвертой главе проведен сравнительный анализ алгоритмов федеративного обучения FedAvg, FedOpt и FedProx для прогнозирования психоэмоционального выгорания студентов. Показано, что FedAvg ограниченно эффективен при non-IID данных, тогда как FedOpt обеспечивает более быструю и устойчивую сходимость за счет серверной оптимизации, а FedProx повышает стабильность обучения посредством проксимальной регуляризации, но с более медленной сходимостью. В результате FedOpt выбран в качестве основного алгоритма, FedProx как устойчивое решение при высокой гетерогенности данных, а FedAvg используется в качестве базового эталона.

Пятая глава посвящена описанию веб-платформы для сбора данных о питании, физической активности, сне и психоэмоциональном выгорании студентов, а также анализу структуры и адаптации опросника Maslach Burnout Inventory (MBI). Рассмотрены методы обработки non-IID данных, обеспечения конфиденциальности и алгоритмы прогнозирования уровня выгорания. Описаны подходы к кодированию данных, автоматизированному анализу и выявлению факторов, влияющих на психоэмоциональное состояние, а также обозначены ограничения исследования и направления дальнейшего развития.

В главе «Заключение» показано, как реализована и исследована федеративная архитектура обучения для анализа психоэмоционального состояния студентов на распределенных конфиденциальных данных.

Рассмотрено, что адаптация алгоритмов FedAvg, FedOpt и FedProx к недифференцируемой модели Random Forest Regression обеспечивает устойчивую сходимость глобальной модели в условиях non-IID данных. Полученные результаты подтверждают практическую применимость федеративного обучения в образовательной среде при сохранении конфиденциальности информации.

Личный вклад автора. Все основные результаты, описанные в диссертации, выполнены и собраны автором. Кроме того, основные результаты исследований, анализы, модели, программы созданы автором, выводы сделаны на основе результатов, полученных от работы и исследования PhD докторанта.

Структура и обьем работы. Диссертация включает введение, 5 основных глав, заключения и списка использованных источников. Полный обьем диссертации составляет 144 страницы, включая 46 иллюстраций и 14 таблиц. Список литературы содержит 105 наименований.

1 ОБЗОР И АНАЛИЗ ОСНОВ ФЕДЕРАТИВНОГО МАШИННОГО ОБУЧЕНИЯ

1.1 Теоретические основы федеративных моделей и методов машинного обучения

Построение модели машинного обучения осуществляется за счет совместного участия множества клиентских устройств при обучении федеративного обучения (ФО), но исходные данные не покидают устройства клиентов, что обеспечивает высокий уровень конфиденциальности и защиты персональной информации, при этом для формирования глобальной модели обьединяются результаты обучения локальной модели [1].

Активное развитие технологий искусственного интеллекта (ИИ) [2] и машинного обучения (МО) сопровождается быстрым ростом объемов данных, формируемых современными информационными системами. Значительная часть этих данных поступает от различных цифровых устройств, сенсоров и онлайн-платформ и представляет ценность для аналитической обработки, которые содержат персональную и чувствительную информацию, что существенно усложняет задачи обеспечения информационной безопасности и защиты приватной жизни пользователей.

Традиционные подходы к машинному обучению основываются на централизованных архитектурах [3], в которых данные от множества клиентов или узлов собираются и хранятся на одном сервере для обучения модели, несмотря на высокую вычислительную эффективность и возможность использования мощных аналитических инструментов, но существует ряд существенных ограничений, такие как концентрация данных в одном месте значительно увеличивает риск их утечки или несанкционированного доступа, централизованная обработка данных требует значительных ресурсов для хранения и вычислений, что усложняет масштабирование системы, юридические ограничения и нормативные требования накладывают строгие ограничения на передачу и хранение персональной информации, что делает централизованный сбор данных практически неосуществимым для ряда критически важных сфер.

Термин «федеративное обучение» начал активно использоваться в научных публикациях с 2015 года. В рамках обзора научной литературы, представленной в международных библиографических и наукометрических базах данных Scopus, IEEE, Google Scholar и других профильных ресурсах, была выполнена систематическая оценка состояния и тенденций развития данного научного направления. Согласно проведенному литературному обзору, в 2015 году [4] и 2016 году [5] появились первые исследования, связанные с применением федеративного усреднения в телекоммуникационных системах. В последующие годы одной из ключевых исследовательских задач стало снижение коммуникативной нагрузки в процессах федеративного обучения. В 2017 году [6] и 2018 году [7] научное сообщество сосредоточилось на разработке стратегий распределения ресурсов

и оптимизации алгоритмов, позволяющих снизить затраты на передачу данных между узлами сети, кроме того, в этот период активно изучались механизмы обеспечения устойчивости к дифференциальным атакам на конфиденциальность [8].

Для сохранения конфиденциальности и безопасности при обработке данных необходимо разработать такие методы, которые осуществляют обучение моделей прямо на локальных устройствах без необходимости передачи исходных данных на центральный сервер [9], что будет решать проблему безопасности.

В условиях современной цифровой среды одной из приоритетных задач становится создание методов анализа и прогнозирования данных, способных эффективно функционировать при строгих требованиях к защите и конфиденциальности информации. Разработка и совершенствование алгоритмов федеративного обучения, ориентированных на работу с гетерогенными и распределенными источниками данных, является значимым направлением научных исследований, способствующим одновременно развитию аналитических технологий и повышению уровня информационной безопасности в различных отраслях.

Обладая высокой эффективностью, централизованное обучение сталкивается с рядом ключевых проблем: передача и хранение чувствительной информации пользователей на центральном сервере увеличивает риск утечки данных и несанкционированного доступа, что делает вопрос конфиденциальности одной из главных проблем; процесс централизованного обучения требует значительных вычислительных мощностей, что затрудняет его масштабирование при увеличении объемов данных и ведет к высоким затратам на обработку информации; передача больших объемов данных между клиентскими устройствами и сервером может перегружает сеть и замедляет обработку информации и препятствует процессу работы обучения в реальном времени. Следовательно, возникает необходимость разработки таких методов, как федеративное обучение, которое дает возможность обрабатывать данные децентрализованно и уменьшать утечку информации, нагрузку на вычислительные мощности и перегрузку сети.

Для решения этих проблем был разработан децентрализованный подход ФО, который позволяет обучать модели непосредственно на клиентских устройствах, сохраняя локальность исходных данных. Вместо передачи всех наборов данных в центральный узел, ФО позволяет каждому клиенту обучать модель локально и отправлять обновления модели (например, весовые параметры или градиенты) только на центральный сервер [10], где они объединяются для обучения глобальной модели.

Метод, основанный на принципах ФО [11], обладает рядом существенных преимуществ, таких как повышенная защита конфиденциальности, так как данные пользователей остаются на их клиентских устройствах, обработка данных происходит непосредственно на локальных устройствах, что позволяет обеспечить соответствие нормативным

требованиям по защите персональной информации, таким как GDPR [12] и HIPAA. Исходные данные никогда не покидают устройство пользователя и это особенно важно в критически значимых сферах, где конфиденциальность является приоритетной, включая здравоохранение, финансы и образование. В традиционных централизованных системах обработка больших объемов данных требует значительных вычислительных ресурсов, что может приводить к перегрузке серверной инфраструктуры и снижению общей производительности системы. В отличие от централизованного подхода, ФО распределяет вычислительные задачи между клиентскими устройствами, позволяя выполнять локальное обучение и передавать на сервер только агрегированные обновления модели, что существенно снижает нагрузку на центральную инфраструктуру, повышает эффективность и улучшает быстродействие системы.

При централизованной передачи больших данных требует значительных сетевых ресурсов, прегрузке сети и заметному росту задержек. ФО решает проблему тем, что вместо самих данных передаются обновления параметров модели, уменьшая обьем передаваемой информации, снижая требования к сетевым ресурсам и позволяя обновлять глобальную модель. Масштабируемость является значительным преимуществом ФО, таким, что распределенная архитектура обрабатывает данные, которые поступают с множества клиентских устройств, работающих в различных средах [13] и делает его многообещающим для применения в Интернете вещей (IoT), мобильных и облачных вычислениях, при необходимости одновременной обработки данных полученных от большого количества источников. ФО позволяет моделям масштабироваться и подстраиваться к меняющимся условиям эксплуатации для устойчивости и эффективности информационных систем. Применение ФО повышает уровень защиты конфиденциальности и безопасности данных, оптимизирует вычислительные ресурсы, снижает наградку на сеть и обеспечивает масштабируемость решений.

1.2 Ключевые компоненты федеративного обучения

Федеративное обучение состоит из следующих основных компонентов: Клиентские устройства представляют собой множество распределенных вычислительных узлов, таких как смартфоны, планшеты, IoT-устройства, серверы и другие устройства, обладающими вычислительными ресурсами. Каждый из этих узлов обладает способностью обрабатывать локальные данные, что позволяет выполнять обучение моели прямо на устройстве, не передавая исходные данные на центральный сервер. При локальном обучении на каждом клиентском устройстве происходит процесс локального обучения модели, алгоритмы машинного обучения применяются непосредственно к локальному набору данных, что позволяет модели адаптироваться к специфике данных конкретного пользователя или устройства, что обеспечивает персонализацию и высокую релевантность обученной модели.

Процесс вычисления обновлений модели: на клиентские устройствах выполняется локальное обучение исходных данных и вычисляются обновления модели, такие как градиенты или изменения весовых коэффициентов. Обновления представляют собой сжатую информацию о необходимых корректировках глобальной модели, предоставляя возможность адаптировать ее без необходимости передачи необработанных данных на центральный сервер, что выполняет условия конфиденциальности. Благодаря локальному обучению персональные и и чувствительные данные остаются на стороне клиента, что соответствует нормативным требованиям по защите данных и гарантирует высокий уровень информационной безопасности. Такой подход особенно актуален для жизненно важных сфер, таких как здравоохранение, финансы, образование и т.д.

Разнообразие устройств и гетерогенность являются важными фвкторами, влияющими на процесс ФО. Клиентские устройства могут значительно различаться по вычислительной мощности, обьему доступных данных и уровню сетевго подключения, что приводит к вариативности в процессе обучения. Такие различия создают дополнительные сложности при синхронизации обновлений и требуют разработки адаптивных алгоритмов, способных учитывать специфику каждого устройства.

Клиентские устройства играют ключевую роль в ФО, обеспечивая локальную обработку данных, вычисление обновлений модели и сохранение конфиденциальности информации, что является основой для постороения эффективных распределенных систем МО.

В процессе ФО агрегация обновлений модели является основным этапом, во время которого центральный сервер собирает и обьединяет обновления, обработанные на локальных устройствах. После завершения локального обучения клиенты передают на сервер параметры модели, такие как градиенты или изменения весовых коэффициентов, исключая передачу самих данных, что обеспечивает кониденциальность. Сервер агрегирует полученные обновления, применяя различные алгоритмы усреднения и оптимизации, включая FedAvg [14], которое выполняет средневзвешенное обьединение обновлений, а также, которые учитывают дополнительные параметры, обеспечивая устойчивость и адаптивность глобальной модели к различным условиям распределенных вычислений.

Центральный сервер выолняет основную роль в системе ФО, собирая обновлени модели, вычисленные на клиентских устройствах. После завершения локального обучения при каждом клиенте, устройства отправляют только вычиленные изменения, например, градиенты или изменения весовых коэффициентов на сервер, при этом исходные данные остаются на устройствах, что обеспечивает высокий уровень конфиденциальности и позволяет минимизировать риски утечки информации [15]. Сервер агрегирует эти обновления с учетом данных каждого клиента. Агрегация глобальной модели является завершающим этапом агрегации в ФО, при этом центральный сервер обновляет параметры глобальной модели на основе агрегированных

данных, полученных от клиентов. Интеграция обновлений, собранных с различных устройств, позволяет модели учитывать разнообразие локальных данных, что особенно важно при гетерогенных распределений [16], что способствет повышению точности и устойчивости модели, при этом адаптируя к специфическим особенностям каждого клиента. Для достижения оптимальной сходимости используются методы адаптивного обновления, включая FedAvg [17], а также более сложные алгоритмы оптимизации FedOpt и FedProx, которые учитывают различия в вычислительных мощностях и характеристиках данных клиента. В итоге обученная глобальная модель становится более обобщенной и применимой к широкому спектру данных, обеспечивая высокую эффективность при последующих итерациях локального обучения [18].

Заключительным этапом итерационного процесса ФО является перераспределение обновленной модели. После обновления параметров на сервере глобальная модель отправляется обратно клиентам и далее используется в следующем цикле локального обучения. Процесс двустороннего обмена информацией между сервером и клиентами повторяется в течение нескольких раундов обучения, что позволяет модели пошагово улучшаться и адаптироваться к новым данным. По мере накопления обновлений и их интеграции в глобальную модель повышается ее точность, устойчивость и способность к обобщению, что особенно важно при работе с неоднородными данными [19], что обеспечивает динамическую адаптацию модели к изменениям в локальных данных пользователей, а также учитывает особенности вычислительных возможностей каждого клиента, делая ФО более эффективным и масштабируемым.

Основную роль в организации ФО играют управление и координация процесса. Помимо выполнения агрегации и обновления глобальной модели, федеративный сервер отвечает за синхронизацию работы клиентских устройств, мониторинг прогресса обучения и адаптацию параметров [20], что включает в себя управление распределением вычислительных нагрузок, определение критериев завершения обучения, а также динамическую настройку гиперпараметров для оптимизации сходимости модели.

Контроль синхронизации позволяет минимизировать задержки в передаче обновлений, обеспечивая согласованность обучения между клиентами. Мониторинг прогресса обучения включает анализ изменений в глобальной модели, оценку качества сходимости и выявление потенциальных проблем, таких как, небалансированное обновление весов или расхождение локальных моделей. Сервер выполняет адаптацию праметров обучения, регулируя частоту обновлений, количество вовлекаемых клиентов и стратегию агрегацию, повышая эффективность ФО, оптимизировать использование сетевых и вычислительных ресурсов [21-23], а также гарантировать устойчивость модели в условиях распределенного и гетерогенного обучения.

Федеративный сервер [24] является центральным звеном системы, которое не только обьединяет обновления модели, полученных с

распределенных устройств, но и гарантирует, что итоговая глобальная модель соответствует высоким требованиям по точности, масштабируемости и безопасности данных.

В процессе ФО обеспечение безопасности, конфиденциальности и целостности передачи данных между клиентами и центральным сервером является ключевым аспектом, используя специализированные защищенные коммуникационные протоколы, которые позволяют безопасно обмениваться обновлениями модели без раскрытия исходных данных [25].

Основные приемы защищенного обмена в ФО включают несколько аспектов. Конфиденциальность достигается за счет того, что данные остаются на клиентских устройствах, а передаваемые обновления шифруются перед отправлением на сервер, что минимизирует риски утечки и несанкционированного доступа. Целостность обеспечивается проверкой корректности и неизменности переданных параметров модели, что предотвращает возможные атаки злоумышленников на процесс обучения. Аутентификация играет важную роль в ФО, так как позволяет использовать механизмы верификации клиентов, обеспечивая доступ к обучению только доверенным участникам [26].

ФО представляет собой децентрализованный подход к МО, при котором модели обучаются непосредственно на локальных устройствах, а не на централизованном сервере. Метод разрабатывался с целью решения основных моментов современной обработки данных, особенно связанных с защитой конфиденциальности, экономией ресурсов и снижением затрат на передачу информации [27, 28]. Основное преимущество ФО заключается в том, что данные пользователей не покидают их устройства. Вместо этого происходит вычисление обновлений модели прямо на клиентских устройствах, а затем эти обновления передаются на центральный сервер, где они агрегируются для формирования глобальной модели [29]. соответственно, исходные данные остаются локальными, а передаются лишь сводные параметры, что также отвечает современным нормативным требованиям по защите персональных данных [30].

Эффективность ФО напрямую зависит от среды его функционирования. В первую очередь, это аппаратное обеспечение клиентских устройств, которое должно обеспечивать достаточные вычислительные мощности для проведения локального обучения. Программное обеспечение, реализующее алгоритмы ФО, должно быть адаптировано для работы в условиях ограниченных ресурсов и разнородности устройств. Сетевая инфраструктура играет важную роль, так как качество и скорость передачи обновлений модели напрямую влияют на время сходимости глобальной модели и общую эффективность системы.

Области применения ФО варьируются от медицины и финансов до образования и военного образования. Каждая из этих отраслей предъявляет специфические требования к безопасности и конфиденциальности обрабатываемых данных, а также к точности и адаптивности обучаемых

моделей. Например, в здравоохранении федеративное обучение позволяет создавать персонализированные модели диагностики, не нарушая требований к защите медицинской информации [31], в финансовом секторе - разрабатывать системы обнаружения мошенничества, а в образовании - обеспечивать адаптивное обучение, учитывающее индивидуальные особенности студентов [32]. Федеративное обучение не только решает проблему конфиденциальности за счет локальной обработки данных, но и оптимизирует использование вычислительных и сетевых ресурсов, что делает его особенно перспективным для внедрения в условиях современной цифровой трансформации различных отраслей экономики [33].

1.3 Меры безопасности и конфиденциальности

ФО обеспечивает конфиденциальность данных, сохраняя необработанную информацию на локальных устройствах, но для увеличения уровня безопасности применяются дополнительные меры защиты. Дифференциальная конфиденциальность добавляет шум к обновлениям модели перед их передачей на сервер, что предотвращает возможность восстановления исходных клиентских данных.

Безопасные многосторонние вычисления Secure Multi-Party Computation (далее SMPS) позволяют зашифровать обновления модели таким образом, чтобы центральный сервер не имел доступа к отдельным клиентским данным, но при этом может агрегировать их [34]. Гомоморфное шифрование предоставляет возможность выполнять вычисления непосредственно над зашифрованными данными, не требуя их предварительной расшифровки, что дополнительно снижает риски утечки информации. ФО обеспечивает конфиденциальность данных, сохраняя необработанные данные на локальных устройствах, однако дополнительные меры безопасности позволяют значительно повысить уровень защиты и надежности процесса обучения. Дифференциальная конфиденциальность предотвращает возможность восстановление исходных данных за счет добавления случайного шума к обновлениям модели перед их передачей не сервер, что обеспечивает защиту даже случае анализа переданных параметров злоумышленниками [35]. Для обеспечения персонализированного обучения при сохранении конфиденциальности данных предлагается метод персонализированное ФО, основанное на использование безопасных многосторонних вычислений SMPC [36], которое представляет собой криптографический метод, позволяющий зашифровать обновления модели таким образом, чтобы центральный сервер не получает доступ к отдельным колиентским данным, но при этом корректноагрегирует параметры модели.

Гомоморфное шифрование явдяется передовой технологией, позволяющей выполнять вычисления над зашифрованными данными без их расшифровки, что значительно повышает уровень конфиденциальности. Дополнительно, устойчивость к византийским ошибкам (BFT) защищает процесс обучения от вредоносным клиентов, которые могут отправлять

искаженные обновления, пытаясь отравить модель и снизить ее точность. В статье [37] представлен метод децентрализованного ФО, устойчивого к византийским сбоям. Авторы предлагают одноранговую сеть (P2P), где транспортное средство сотрудничает в обучении модели без использования центрального сервера, а механизмы обеспечивают защиту от ненадежных или вредоносных участников. В совокупности эти методы делают ФО безопасным и надежным решением для критически важных областей, как здравоохранение, финансы и государственные информационные системы. Чаще ФО применяется в различных областях [38], где конфиденциальность, безопасность данных и децентрализованное обучение имеют решающее значение, как сфера здравоохранения и медицинские исследования, где ФО используется для прогнозирования и диагностики заболеваний [39], позволяя больницам и научно-исследовательским институтам обучать модели МО на децентрализованных данных пациентов без передачи конфиденциальной информации. В области медицинской визуализации ФО способствует совместному обучению моделей в разных больницах, что повышает точность диагностики, например, в выявлении онкологических заболеваний и анализе рентгенологических снимков [40-42], кроме того, носимые устройства, как умные часы и фитнес-трекеры применяют ФО для персонализированных рекомендаций по здоровью, при этом обеспечивая конфиденциальность пользователей. В финансовой сфере и банковском деле ФО используется для обнаружения мошенничества, позволяя банкам совместно обучать модели выявляя подозрительных транзакций без раскрытия данных клиентов [43], и в кредитном скоринге несколько финансовых организаций могут использовать ФО для улучшения моделей оценки рисков, не обьединяя финансовые записи клиентов в единую базу. Метод активно применяется для борьбы с отмыванием денег, помогая выявлять подозрительные транзакции в различных финансовых организациях при соблюдении требований конфиденциальности. В области умных городов в Интернете вещей [44], ФО применяется для прогнозирования и оптимизации дорожного движения, позволяя городской инфраструктуре анализировать потоки трафика, используя децентрализованные данные от транспортных средств и датчиков. ФО способствует управлению энергопотреблением, позволяя интеллектуальным энергосетям оптимизировать модели потребления электроэнергии, не подвергая риску данные пользователей. В сфере автономного транспорта подключенные автомобили используют ФО для обмена информацией о дорожных условиях и схемах вождения, сохраняя при этом конфиденциальность данных водителей. В области периферийных вычислений и мобильных приложений ФО применяется для персонализированных рекомендаций, улучшая работу искусственного интеллекта в таких приложениях, как клавиатуры, например Gboard от Google и виртуальные помощники. В моделях речи и языка ФО помогает обучать модели распознавания речи без загрузки голосовых данных на центральные серверы, что повышает уровень защиты информации. Федеративное периферийное обучение снижает зависимость от облачных

вычислений, обучая модели непосредственно на распределенных периферийных устройствах, что уменьшает задержки и затраты на полосу пропускания. ФО активно используется в сфере кибербезопасности и обнаружения угроз. В системах обнаружениях вторжений метод позволяет выявлять угрозы кибербезопасности в реальном времени, анализируя данные от нескольких организаций [45]. В обнаружении вредоносных программ ФО обеспечивает совместную работу распределенных устройств [46] для выявления новых угроз без раскрытия журналов безопасности. В обнаружении ботнетов ФО способствует разработке эффективных стратегий борьбы с кибератаками, используя данные из разных сетей для своевременного реагирования. В области образования и электронного обучения ФО используется для создания персонализированных моделей обучения, позволяя университетам и онлайн-платформам разрабатывать адаптивные учебные программы на основе децентрализованных данных о студентах. Метод применяется на разработке интеллектуальных преподавателей и чат-ботов, которые повышают точность своих ответов, анализируя взаимодействие с учащимися. В сфере контроля честности экзаменов ФО улучшает обнаружение случаев мошенничества с помощью ИИ при проведении онлайн-тестирования, при этом не нарушая конфиденциальность студентов. ФО становится ключевой технологией для множества отраслей, обеспечивая баланс между эффективностью моделей и защитой конфиденциальных данных.

ФО хорошо подходит для вышеупомянутых областей применения [47], поскольку оно обеспечивает конфиденциальность и безопасность данных, сохраняя пользовательскую информацию на локальных устройствах и исключая необходимость ее передачи на централизованный сервер, что снижает потребность в масштабной централизованной инфраструктуре для хранения данных, уменьшает затраты на ее обслуживание. ФО позволяет обучать и обновлять модели в режиме реального времени, обеспечивая их адаптацию без прерывания работы системы. Одним из существенных преимуществ является поддержка обучения на основе гетерогенных данных, что делает этот подход эффективным в разнообразных и неоднородных средах.

Среда ФО включает децентрализованные клиентские устройства, защищенные каналы связи и специализированные фреймворки ФО, обеспечивающие процесс распределенного обучения с сохранением конфиденциальности данных. Области применения ФО охватывают здравоохранение, финансы, Интернет вещей, мобильные приложения, кибербезопасность и образование отрасли, в которых защита персональной информации, децентрализованный интеллект и обработка данных в реальном времени имеют решающее значение [48]. Совокупность указанных факторов делает ФО перспективным подходом для развития интеллектуальных систем на основе ИИ, отвечающих современным технологическим, этическим и нормативным требованиям [49].

1.4 Требования к развитию и совершенствованию ФО

Развитие ФО как парадигмы распределенного МО требует постоянного совершенствования для повышения его эффективности, масштабируемости, безопасности и применимости в различных сферах, что для обеспечения его эволюции необходимо учитывать несколько важных требований, таких как масштабируемость и вычислительная эффективность, поскольку ФО должно адаптироваться к работе в больших сетях клиентских устройств, сохраняя при этом высокую точность и минимальные затраты на вычисления. Для достижения этой цели важно внедрение эффективных методов агрегации моделей, таких как федеративное усреднение, которые должны быть оптимизированы для снижения накладных расходов на передачу данных без ущерба для точности модели. Системы ФО должны поддерживать адаптивное участие клиентов, что предполагает динамический выбор устройств для обучения в зависимости от их вычислительных возможностей, пропускной способности сети и качества локальных данных.

Важную роль играют стратегии обучения с учетом ресурсов, которые направлены на разработку облегченных моделей, подходящих для маломощных устройств, таких как системы Интернета вещей, мобильные устройства и периферийные вычисления и расширяет применение ФО в средах с ограниченными вычислительными ресурсами, сохраняя при этом эффективность обучения и обеспечивая возможность масштабирования распределенных моделей [50].

Развитие и совершенствование федеративного обучения требуют многогранного подхода, который учитывает масштабируемость, конфиденциальность, эффективность коммуникации, персонализацию, стандартизацию, сходимость моделей и устойчивость к неоднородным данным. Будущие направления исследований должны быть сосредоточены на разработке систем ФО, которые балансируют точность с вычислительной эффективностью, обеспечивают механизмы сохранения конфиденциальности и обеспечивают адаптируемость к различным реальным приложениям. Соответствуя этим требованиям, ФО может быть широко принят в таких отраслях, как здравоохранение, финансы и интеллектуальные периферийные вычисления, при этом соблюдая этические и нормативные требования.

Оптимизация играет важную роль в ФО, поскольку децентрализованный характер подхода создает ряд трудностей такие как, эффективность связи, сходимость моделей и неоднородность клиентских данных. В этом разделе представлен обзор общих методов оптимизации ФО и конкретные методы, которые разработаны различные методы оптимизации для повышения производительности, стабильности и точности в средах ФО. В ФО для достижения высокого качества модели разработали стратегии оптимизации, которые ускоряют сходимость моделей, справляются с неоднородностью данных плюс снижают вычислительные и сетевые затраты.

Снижение затрат на передачу данных между клиентами и центральным сервером является основной задачей ФО. Высокая частота обмена

параметрами приводит к задержкам и перегрузке сети в условиях большого количества клиентов и ограниченной пропускной способности. Методы, направленные на сокращение обьема передаваемых данных и повышение продуктивности распределенного обучения [51].

Обеспечение сходимости обьединенных моделей в условиях распределенного обучения требует применения методов оптимизации, которые учитывают ограничения вычислительных ресурсов, гетерогенность данных и нестабильность обновлений со стороны клиентов. Одним их методов является адаптивная федеративная оптимизация (далее FedOpt), использующая адаптивные методы обновления параметров на сервере, например, Adam или Adagrad, для ускорени сходимости и снижения влияния разнородных данных на обучение [52].

Другой подход - Federated Proximal (далее FedProx), который вводит проксимальный член в функцию потерь, ограничивая расхождение локальных обновлений между клиентами и глобальной модели, что является полезным в ситуациях, где данные клиентов не являются гетерогенными non-IID, что может приводить к нестабильности обучения, FedProx помогает сглаживать обновления модели, обеспечивая более плавную и стабильную сходимость, даже если клиенты имеют разные вычислительные возможности и обьемы данных. Плюсом при применении методов будет повышение продуктивности ФО, что улучшает точность, стабильность и скорость сходимости модели в условиях ограниченных вычислительных ресурсов и высокой гетерогенности данных.

Архитектурные модели федеративного обучения. Типы ФО.

В ФО используются различные методы обучения, которые входят в структуру ФО, как показано на рисунке 1.1, который включает в себя синхронный, асинхронный и полусинхронный подходы. Каждый из них имеет свои особенности и применяется в зависимости от требований к успешности обучения, стабильности обновлений и адаптации модели к гетерогенным данным. Синхронное обучение обеспечивает высокую точность и последовательность обновлений, но может замедляться из-за медленных клиентов. Асинхронное обучение ускоряет процесс, позволяя клиентам обновлять модель независимо, однако может привести к расхождению параметров. Полусинхронный подход является альтернативным решением, позволяя учитывать обновления от части клиентов перед агрегацией, что снижает задержки и улучшает адаптацию модели в условиях неоднородного распределения данных

В ФО стратегии синхронизации и агрегации задают порядок взаимодействия клиентских устройств с сервером при формировании глобальной модели. Они определяют моменты передачи локальных обновлений и способы их обьединения, обеспечивая согласованность процесса распределенного обучения.

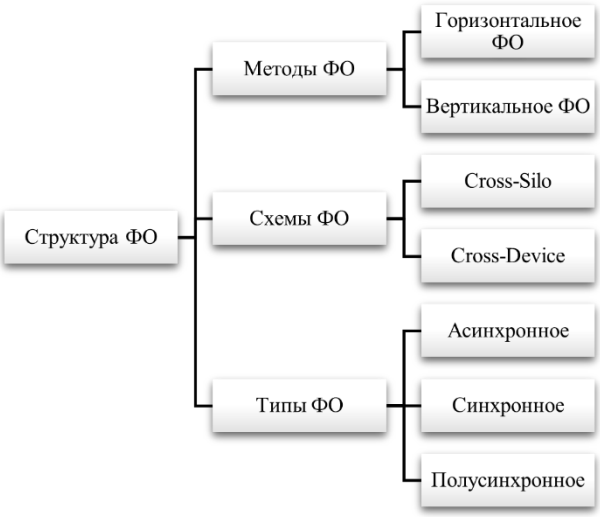


Рисунок 1.1 - Структура ФО

В рамках данного исследования использовано синхронное ФО, при котором все выбранные клиенты завершают локальное обучение в пределах одного раунда, а сервер агрегирует их обновления параллельно, обеспечивая согласованность глобальной модели и устойчивую сходимость, что особенно важно при работе с чувствительными и гетерогенными данными о психоэмоциональном состоянии студентов.

Асинхронные и полусинхронные стратегии, несмотря на снижение задержек, могут приводит к рассогласованности обновлений и нестабильности модели, поэтому они не применялись в условиях ограниченного числа клиентов и контролируемой архитектуры [53]. Выбор синхронной стратегии агрегации в рамках исследования обеспечил стабильность процесса обучения, корректность федеративной агрегации и надежную работу глобальной модели в условиях non-IID распределения данных. При использовании синхронного подхода достигается высокую согласованность обновлений модели, что является решающим фактором для повышения точности и устойчивости прогнозирования при анализе чувствительных данных психоэмоционального состояния студентов.

Синхронное ФО предполагает, что глобальная модель обновляется только после завершения локального обучения всеми клиентами и получения их обновлений, что обеспечивает последовательность обновлений и способствует высокой точности и стабильности итоговой модели, но одним из его недостатков является чувствительность к медленным клиентам, которые могут замедлять процесс агрегации из-за разной вычислительной мощности или нестабильного сетевого соединения.

В асинхронном федеративном обучении клиенты обновляют глобальную модель независимо, как только завершают локальное обучение, однако сокращает время ожидания и повышает скорость обучения, особенно в распределенных и нестабильных средах с различными вычислительными мощностями клиентов, но отсутствие единого момента агрегации может

привести к расхождению моделей, если обновления поступают с разными интервалами [54].

В исследовании требовалась стабильность динамики обучения и контролируемые обновления, поэтому асинхронный ФО не был выбран. Его применение могло бы привести к дисбалансу в обучении, где обновления от быстрых клиентов оказывали бы непропорционально большое влияние на итоговую модель.

Полусинхронное федеративное обучение представляет собой гибридный подход, который комбинирует элементы синхронного и асинхронного методов. В этом методе глобальная модель обновляется после получения обновлений от определенного подмножества клиентов, например, 80% участников, что позволяет уменьшить задержки, вызванные медленной работой отдельных клиентов, при этом сохраняя согласованность обновлений, хотя полусинхронный подход мог бы предложить гибкость в обучении, но все-таки не обеспечивал полной согласованности модели. В рамках данного исследования учитывалось полное участие всех клиентов, что позволило избежать потенциальных проблем со сходимостью и сохранить стабильность обновлений [55]. Синхронное обучение выбрано в качестве основного метода, поскольку обеспечивает стабильность, равномерность обновлений и согласованность модели, а также гарантирует равномерный вклад всех клиентских моделей в процесс федеративной агрегации, обеспечивая последовательные и контролируемые обновления, что важно в условиях гетерогенных данных, где асинхронное обновление могло бы привести к сильному расхождению моделей. Асинхронный и полусинхронный подходы не были применены, так как их недостатки (расхождение моделей, неконтролируемые обновления, влияние быстрых клиентов на итоговую модель) могли негативно повлиять на качество обучения в условиях гетерогенных данных.

Схемы ФО и их применение. Горизонтальное и вертикальное ФО.

Федеративное обучение классифицируется на основе того, как данные распределяются между клиентами. Два основных типа - это горизонтальное федеративное обучение (далее ГВО) и вертикальное федеративное обучение (далее ВФО), каждый из которых предназначен для разных сценариев.

ГВО применяется в случаях, когда различные пользователи или устройства обладают данными с одинаковым пространством признаков, но из разных выборок, где каждый клиент обучает модель локально, используя свои данные, которые имеют идентичную структуру признаков, но представляют разные наблюдения и после локального обучения обновления модели передаются на центральный сервер, где они агрегируются для формирования единой глобальной модели, что показывает метод обучения особенно полезен в ситуациях, когда организации или устройства работают с аналогичными типами данных, но не могут передавать их напрямую из-за требований к конфиденциальности или распределенной инфраструктуры. Примером применения ГВО является совместное обучение моделей в медицинских

учреждениях, где несколько больниц могут участвовать в создании общей модели диагностики заболеваний, используя данные своих пациентов, при этом каждое медицинское учреждение работает с одним и тем же набором признаков, такими как возраст, симптомы и диагнозы, но с разными пациентами, позволяя моделям учиться на более обширных и разнообразных данных, сохраняя при этом конфиденциальность медицинской информации, так как исходные данные не передаются за пределы конкретной больницы.

В рамках исследования ГВО было выбрано в качестве основной стратегии федеративного обучения, поскольку клиентские устройства обладают одинаковым пространством признаков, включающим параметры продолжительности сна, уровня физической активности, режимов питания и эмоционального благополучия, при этом каждое учреждение обучает локальную модель на индивидуальном наборе данных, а затем передает обновления модели на сервер, где они агрегируются для формирования глобальной модели, индивидуальные данные студентов учитываются, одновременно обеспечивая конфиденциальность информации, так как исходные данные остаются на локальных устройствах. Централизованная агрегация обновлений модели позволяет повысить точность и обобщающую способность модели, что делает данный метод особенно эффективным для анализа данных, связанных с психологическим и физическим состоянием студентов в условиях распределенного машинного обучения [56].

Горизонтальное федеративное обучение направлено на горизонтальную централизацию данных похожих устройств в пределах одного уровня или горизонта, например, общие метки классов и примеры данных могут объединяться горизонтально [57]. Обмен данных происходит между подобными устройствами, и обновленные модели передаются друг другу. Метод горизонтального федеративного обучения эффективен, когда общие шаблоны или закономерности могут быть извлечены из данных различных устройств, предоставляя возможность устройствам схожего типа обмениваться знаниями и учиться друг у друга, что может быть полезно, когда данные имеют общие характеристики. ГВО [58] представляет собой архитектуру распределенного обучения, при которой каждый клиент обладает собственным набором данных с одинаковым пространством признаков, где модель обучается локально на стороне каждого участника, после чего обновленные параметры передаются на центральный сервер для агрегации с использованием методов взвешенного усреднения. Полученная глобальная модель рассылается обратно клиентам и итеративно уточняется в последующих раундах обучения, учитывая локальные особенности данных и одновременно формируя обобщенную модель, сохраняя конфиденциальность, поскольку исходные данные не покидают клиентские узлы. В контексте исследования каждый клиент оперирует локальными данными опросника MBI, отражающими психоэмоциональное состояние студентов.

Метод позволяет решать технические и практические задачи путем распределения данных на различные устройства. Алгоритм сравнивает

характеристики и связывает их соответствующим образом. Пример горизонтального федеративного обучения представлен, в соответствии с рисунком 1.2.

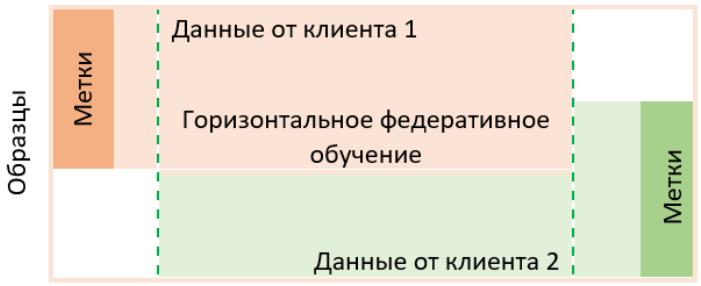


Рисунок 1.2 - Горизонтальное ФО

ВФО применяется в случаях, когда различные организации или участники обладают данными об одних и тех же пользователях, но с различными наборами признаков и в отличие от горизонтального федеративного обучения, где клиентские устройства работают с одинаковыми характеристиками данных, в ВФО объединение информации происходит за счет совмещения разных аспектов данных, принадлежащих одним и тем же субъектам [59]. ВФО особенно полезно в сценариях, когда необходимо объединить информацию из разных источников без раскрытия конфиденциальных данных, например, банк и платформа электронной коммерции могут обслуживать одних и тех же пользователей, но при этом банк располагает данными о финансовых транзакциях, а платформа электронной коммерции собирает информацию о поведенческих предпочтениях покупателей. ВФО позволяет объединить эти данные для создания более точных моделей, например, для предсказания платежеспособности клиента, при этом не передавая сырые данные между организациями [60-62]. Хотелось бы отметить, что в данном исследовании ВФО не использовалось, так как все клиенты собирают одинаковые типы данных (продолжительность сна, физическая активность, приемы пищи и уровень эмоционального благополучия) и каждый клиент обладает полным пространством признаков, необходимость в объединении разнородных данных об одних и тех же пользователях отсутствует, что делает вертикальное федеративное обучение нерелевантным для данной задачи. Вертикальное федеративное обучение ориентировано на объединение вертикально разделенных данных по признакам или атрибутам. Обмен информацией происходит между устройствами, владеющими разными аспектами данных [63]. Модели обучаются на этих частях данных, а затем объединяются для создания общей модели, что дает возможность устройствам с различными характеристиками обмениваться знаниями, что может улучшить качество модели, особенно когда разные устройства содержат уникальные аспекты данных. Например, предположим, у нас есть две организации: одна имеет данные о пациентах,

например, их имена, а другая - медицинскую информацию, например, результаты анализов. Вертикальное федеративное обучение позволило бы этим организациям совместно обучать модель, не раскрывая конкретные имена пациентов и их медицинские результаты [64, 65]. Вертикальное федеративное обучение [66] применяется в сценариях, когда участники располагают различными наборами признаков об одних и тех же объектах. В отличие от горизонтального подхода, вертикальное ФО требует согласования сущностей и использования криптографических протоколов, методов безопасного мультипартийного вычисления и защищенной агрегации, при этом объединяя разнородные атрибуты без раскрытия сырых данных и широко применяется в финансовых и медицинских системах. Высокая вычислительная и коммуникационная сложность вертикального ФО ограничивает его применение в задачах, аналогичных рассматриваемому исследованию. Пример вертикального федеративного обучения представлен, в соответствии с рисунком 1.3.

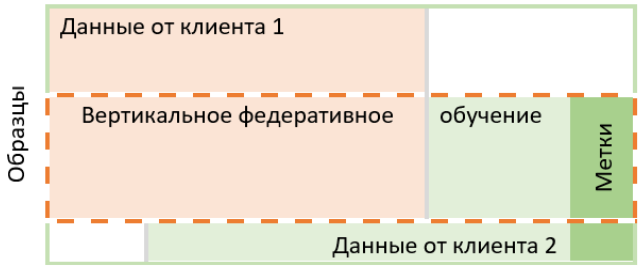


Рисунок 1.3 - Вертикальное ФО

1.5 Типы развертывания ФО. Cross-Silo и кросс-устройство ФО

Федеративное обучение может быть классифицировано в зависимости от количества и типа участвующих клиентов, что определяет особенности вычислительных процессов, требования к сетевой инфраструктуре и устойчивость соединений. В рамках данного исследования был использован подход Cross-Silo федеративного обучения [67]. Известно, что Cross-Silo федеративное обучение применяется в сценариях, где в обучении участвует ограниченное число надежных и устойчивых клиентов, обладающих достаточными вычислительными ресурсами, стабильным сетевым подключением и значительными объемами данных, к таким клиентам относятся, в частности, образовательные учреждения, исследовательские организации и корпоративные системы, а также широко используется в межорганизационных задачах, где требуется высокая точность моделей, контролируемый процесс обучения и централизованная координация агрегации [68, 69].

В рамках нашего исследования клиенты федеративной системы представлены образовательными учреждениями, каждое из которых собирает данные студентов локально и выполняет обучение собственной локальной

модели. Несмотря на то, что данные формируются на персональных устройствах студентов, процесс федеративного обучения организован на уровне учреждений, которые выступают в роли устойчивых и доверенных узлов федеративной архитектуры, что соответствует парадигме Cross-Silo федеративного обучения, где каждый клиент представляет собой агрегированный источник данных с контролируемыми вычислительными ресурсами [70, 71].

Выбор Cross-Silo подхода обусловлен следующими факторами:

- ограниченным и фиксированным числом клиентов;
- стабильным сетевым взаимодействием между клиентами и сервером;
- однородным признаковым пространством данных;
- возможностью синхронной координации процесса обучения.

В исследовании использовано синхронное федеративное обучение, при котором сервер ожидает завершения локального обучения всех клиентов перед выполнением этапа агрегации, что обеспечивает согласованность обновлений, предсказуемую динамику сходимости и устойчивость глобальной модели, что особенно важно при анализе гетерогенных non-IID данных. В работе реализована горизонтальная схема федеративного обучения, поскольку все клиенты оперируют данными с одинаковым пространством признаков, включающим параметры сна, физической активности, питания и психоэмоционального состояния студентов и использование Cross-Silo федеративного обучения в сочетании с синхронной агрегацией и горизонтальной архитектурой позволило обеспечить эффективное распределенное обучение, сохранить конфиденциальность данных и достичь устойчивой сходимости глобальной модели в условиях реальной образовательной среды.

1.6 Виды федеративного обучения

Централизованное федеративное обучение.В централизованном федеративном обучении один центральный сервер отвечает за координацию процесса обучения. Рабочий процесс представлен, в соответствии с рисунком 1.4.

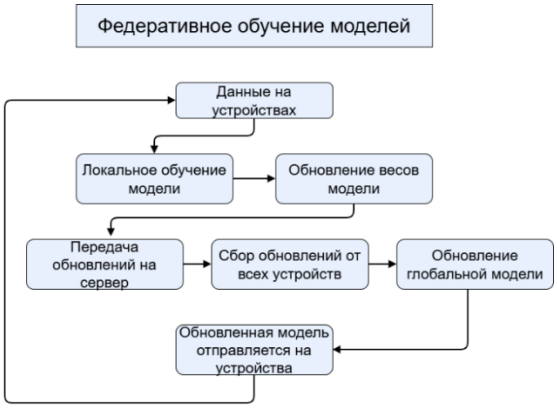


Рисунок 1.4 - Рабочий процесс обучения ФО

Архитектура является наиболее широко используемой и служит основой для многих приложений ФО благодаря своей простоте и хорошо структурированному процессу агрегации. Преимущества и проблемы показаны, в соответствии с рисунком 1.5.

Преимущества	Проблемы
Эффективная синхронизация: обеспечивает согласованные обновления моделей и сокращает расхождения между клиентами. Хорошо зарекомендовавшие себя алгоритмы: совместимы с широко используемыми методами агрегации, такими как FedAvg, FedOpt и FedProx. Проще реализовать: единый координирующий сервер упрощает развертывание.	Проблемы масштабируемости: центральный сервер может стать узким местом по мере увеличения числа клиентов. Единая точка отказа: если центральный сервер скомпрометирован, весь процесс обучения будет нарушен. Риски конфиденциальности: хотя необработанные данные остаются на клиентах, метаданные или обновления моделей по-прежнему могут привести к утечке конфиденциальной информации.

Рисунок 1.5 - Преимущества и проблемы

Децентрализованное федеративное обучение

В децентрализованном федеративном обучении [72] клиенты напрямую общаются друг с другом для обмена обновлениями модели, не полагаясь на центральный сервер. Подход «равный-равному» (P2P) повышает надежность системы и позволяет избежать сбоев в одной точке. Этапы процесса обучения показаны, в соответствии с рисунком 1.6.

Процесс обучения следует следующим основным этапам:
• Каждый клиент обучает локальную модель и делится обновлениями с подмножеством одноранговых узлов вместо центрального агрегатора. • Участники совместно объединяют свои обновления, используя такие методы, как протоколы сплетен или механизмы консенсуса . • Обобщенная модель совершенствуется посредством многократных итераций локального обучения и общения со сверстниками

Рисунок 1.6 - Этапы процесса обучения

Архитектура особенно полезна для сетей, где центральный координирующий сервер нецелесообразен или нежелателен из-за ограничений

конфиденциальности, безопасности или инфраструктуры, где одним из ключевых преимуществ рассматриваемого подхода является повышенная отказоустойчивость, обусловленная отсутствием зависимости от единого сервера, что значительно снижает вероятность сбоя системы и повышает ее надежность, где обеспечивается повышенная конфиденциальность, поскольку клиенты взаимодействуют исключительно с выбранными одноранговыми узлами, что уменьшает риск атак злоумышленников и утечек данных. Еще одним важным преимуществом является масштабируемость, позволяющая эффективно поддерживать крупномасштабные системы с большим количеством клиентов, что делает данный подход применимым в условиях распределенного обучения и анализа данных.

Наряду с преимуществами существует ряд проблем, куда входят повышенная сложность коммуникации, требующая эффективных механизмов одноранговой синхронизации для предотвращения расхождения локальных моделей и корректного обновления глобальной модели и одним из ограничений, является более медленная сходимость, поскольку отсутствие централизованного агрегатора может привести к несогласованности обновлений модели, что увеличивает время достижения оптимального состояния. Метод характеризуется высокими вычислительными затратами, так как клиенты вынуждены не только выполнять локальное обучение модели, но и обеспечивать одноранговую коммуникацию, что требует значительных аппаратных ресурсов, указывая на то, что рассматриваемый подход сочетает в себе как высокую надежность, безопасность и масштабируемость, так и сложность реализации, необходимость точной настройки коммуникационных механизмов и значительные вычислительные затраты.

Иерархическое федеративное обучение.

Иерархическое федеративное обучение (ИФО) [73] представляет собой гибридную архитектуру, в которой агрегация локальных обновлений осуществляется на нескольких уровнях с использованием промежуточных (пограничных) серверов перед финальной агрегацией на центральном сервере. В данной схеме клиенты передают обновления ближайшим пограничным узлам, которые выполняют предварительную агрегацию и направляют результаты на верхний уровень иерархии, что снижает коммуникационную нагрузку на центральный сервер, повышает масштабируемость системы и эффективность обучения в крупномасштабных и георасределенных сценариях. ИФО особенно актуально для областей здравоохранения, умных городов и промышленного IoT, где требуется обработка больших объемов распределенных данных. В свою очередь, иерархическая архитектура характеризуется повышенной сложностью реализации, дополнительными требованиями к синхронизации моделей и усиленными мерами безопасности, поскольку пограничные серверы могут выступать уязвимыми точками системы [74-78]. ИФО обеспечивает эффективную масштабируемость и снижение сетевых затрат, однако его применение требует тщательного проектирования архитектуры и механизмов координации [79]. Выбор данной

архитектуры определяется требованиями приложения к конфиденциальности данных, вычислительным ресурсам и сетевым ограничениям, представленных в таблице 1.1.

Таблица 1.1 - Преимущества и недостатки архитектуры ФО

Архитектура ФО	Преимущества	Недостатки
Централизованное ФО	Эффективная координация, отработанные методы, простота развертывания	Единая точка отказа, проблемы масштабируемости
Децентрализованное ФО	Высокая отказоустойчивость, повышенная конфиденциальность, улучшенная масштабируемость	Сложность коммуникации, более медленная конвергенция
Иерархическое ФО	Сокращение накладных расходов на связь, поддержка крупномасштабных сетей, улучшенная масштабируемость	Высокая сложность системы, потенциальные риски безопасности

Каждый из рассмотренных подходов предполагает определенные компромиссы между вычислительной эффективностью, уровнем конфиденциальности и надежностью системы. Централизованное федеративное обучение остается наиболее распространенной архитектурой благодаря структурированному и контролируемому процессу агрегации. В то же время децентрализованное и иерархическое федеративное обучение представляют собой перспективные альтернативы для крупномасштабных систем и сценариев с повышенными требованиями к конфиденциальности и распределенности вычислений [80].

1.7 Выводы по 1 главе

В первой главе рассмотрены теоретические основы федеративного обучения, его архитектурные модели, режимы синхронизации и области применения. Выбор синхронного федеративного обучения обоснован на основе анализа, обеспечивающего согласованность обновлений и устойчивую сходимость глобальной модели при работе с гетерогенными данными о психоэмоциональном состоянии студентов, учитывая, что все клиенты используют единое пространство признаков, принята горизонтальная схема федеративного обучения в парадигме Cross-Silo, где в роли клиентов выступают образовательные учреждения с локальными наборами данных. Комбинация подходов обеспечивает эффективность агрегации, сохранение конфиденциальности данных и практическую применимость предложенного решения в реальной образовательной среде.

2 КЛАССИФИКАЦИЯ АЛГОРИТМОВ МАШИННОГО И ФЕДЕРАТИВНОГО ОБУЧЕНИЯ

2.1 Алгоритмы машинного обучения

В рамках нашего исследования для прогнозирования психоэмоционального выгорания были протестированы несколько алгоритмов МО, причем каждый обладает своими характеристиками. Классические регрессионные модели линейная регрессия, Lasso и Ridge-регрессия, ансамблевые методы случайный лес, градиентный бустинг, включая XGBoost и CatBoost, а также реккурентные нейронные сети LSTM [81, 82] представлены на рисунке 2.1. Рассмотренные модели дали возможность провести сравнительный анализ различных подходов к прогнозированию. Определение подходящей модели для прогнозирования уровней выгорания, проведя оценку их точности и стабильности.

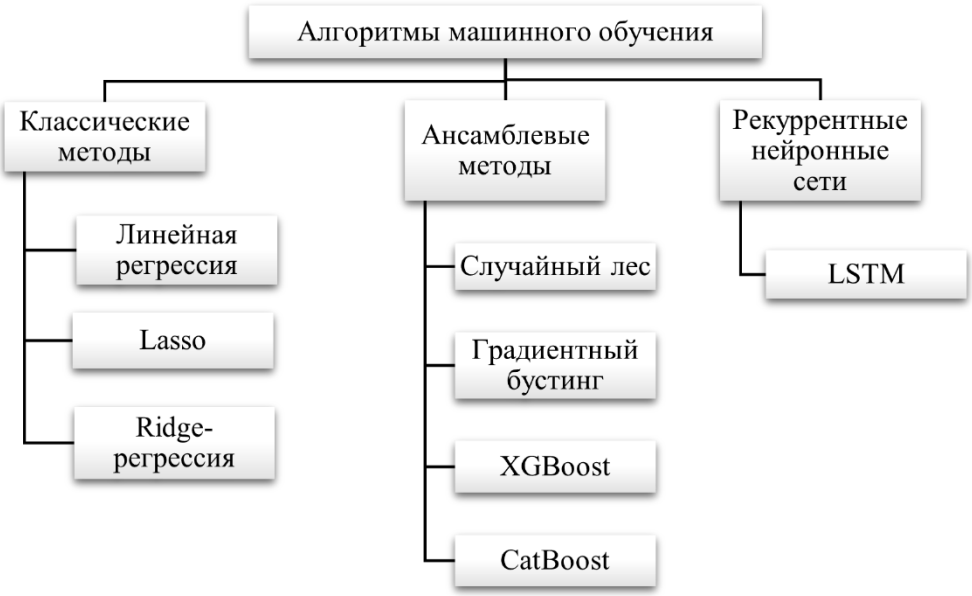


Рисунок 2.1 - Алгоритмы МО

Линейная регрессия, Lasso [83] и Ridge-регрессии [84] основаны на предположении о линейной зависимости между входными признаками и целевой переменной. Методы просты в реализации и легко интерпретируемы, что позволяет оценить влияние каждого признака. Они не способны адекватно моделировать сложные нелинейные взаимосвязи, характерные для психометрических данных, полученных от студентов.

Случайный лес [85], градиентный бустинг, XGBoost и CatBoost [86], представляет собой ансамблевый метод, который строит модель последовательно, корректируя ошибки предыдущих итераций. Алгоритмы демонстрирует высокую точность на сложных данных, при этом есть необходимость настройки гиперпараметров.

Long Short-Term Memory (далее LSTM) - это разновидность рекуррентных нейронных сетей, для работы с временными рядами и выявления длительных зависимостей, несмотря на их потенциал при наличии ярко выраженной временной динамики, в нашем эксперименте данные не обладали очевидной временной структурой плюс обьем выборки был ограничен, что привело к низкой продуктивности модели LSTM.

Модель случайного леса (Random Forest) продемонстрировала оптимальное соотношение точности, устойчивости и вычислительной нагрузки, что было выявлено на основании сравнительного анализа всех рассмотренных алгоритмов, включая метрики Mean Squared Error (MSE), R^2 и т.д. Случайный лес, являясь ансамблевым методом, строит множество решающих деревьев на случайном подмножестве данных и признаков, что позволяет учитывать сложные нелинейные зависимости. Он также отличается высокой устойчивостью к выбросам и шуму. Дополнительным преимуществом является возможность оценки важности признаков, что помогает выявить ключевые факторы, влияющие на психоэмоциональное выгорание, разрабатывая персонализированные стратегии раннего вмешательства. Random Forest был выбран как основной алгоритм, хотя другие алгоритмы также показали высокую точность в определенных аспектах. Его способность моделировать сложные нелинейные зависимости, устойчивость к шуму, относительно невысокая вычислительная сложность и возможность интерпретации важности признаков делают его оптимальным решением для применения в условиях ФО, где данные остаются на клиентских устройствах с ограниченными вычислительными ресурсами.

Классические методы: Линейная регрессия является классическим методом статистического анализа, основанным на предположении о линейной зависимости между независимыми переменными (признаками) и зависимой переменной. Математическая модель линейной регрессии описывается уравнением:

$$\gamma = \beta_0 + \beta_{1x1} + \beta_{2x2} + \cdots + \beta_{nxn} + \varepsilon, \qquad (2.1)$$

где γ – прогнозируемое значение, $x_1, x_2, \cdots, x_n$ – входные признаки, $\beta_0 + \beta_1 + \cdots + \beta_n$ – коэффициенты модели, а ε– случайная ошибка. Преимущества линейной регрессии заключаются в ее простоте, быстроте обучения и высокой интерпретируемости, что позволяет четко понять влияние каждого признака на итоговый прогноз. Метод имеет существенное ограничение и не способен адекватно моделировать сложные нелинейные взаимосвязи, что может быть критичным при прогнозировании психоэмоционального выгорания, где взаимодействия между различными факторами часто являются сложными и многомерными.

Lasso-регрессия (Least Absolute Shrinkage and Selection Operator) представляет собой разновидность линейной регрессии, которая включает *L1* - регуляризацию. Математически модель модифицируется следующим образом:

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(\gamma_i-\beta_0-\sum_{j=1}^{p}\beta_j\,x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}, \quad (2.2)$$

где λ – параметр регуляризации, определяющий степень штрафа за большие значения коэффициентов. Введение *L1* - регуляризации способствует занулению менее значимых коэффициентов, что приводит к автоматическому отбору признаков. При работе с высокоразмерными данными, когда необходимо выделить наиболее значимые детерминанты психоэмоционального выгорания. Несмотря на высокую интерпретируемость, Lasso требует тщательной настройки параметра λ для оптимизации модели.

Ridge-регрессия является аналогом линейной регрессии с *L2* - регуляризацией, которая вводит штраф за квадрат нормы коэффициентов:

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(\gamma_i-\beta_0-\sum_{j=1}^{p}\beta_j\,x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}, \quad (2.3)$$

L2 - регуляризация помогает уменьшить влияние мультиколлинеарности между признаками и стабилизирует оценки коэффициентов, предотвращая переобучение. Ridge-регрессия эффективна при работе с данными, где все признаки вносят вклад в прогнозирование, однако не обеспечивает автоматического отбора признаков, как Lasso.

Ансамблевые методы: Случайный лес (Random Forest) представляет собой ансамблевый метод, основанный на построении большого количества решающих деревьев, каждое из которых обучается на случайном подмножестве данных и признаков. Итоговый прогноз формируется посредством усреднения (для регрессии) или голосования (для классификации) результатов всех деревьев. Метод обладает рядом преимуществ: он способен моделировать сложные нелинейные взаимосвязи, устойчив к выбросам и шуму, а также обеспечивает оценку важности признаков, что особенно полезно при работе с психометрическими данными, такими как результаты опросника MBI. Кроме того, ансамблевый подход случайного леса снижает риск переобучения за счет агрегации множества независимых моделей. Эти свойства делают случайный лес оптимальным выбором для прогнозирования психоэмоционального выгорания, где данные характеризуются высокой разнородностью и сложными взаимосвязями.

Градиентный бустинг (XGBoost, CatBoost) - это семейство алгоритмов, в основе которых лежит построение моделей в виде ансамбля слабых предсказателей, или решающих деревьев, где каждая последующая модель стремится скорректировать ошибки предыдущей. XGBoost (Extreme Gradient Boosting) обладает своей высокой точностью, эффективностью и масштабируемостью, однако требует тщательной настройки гиперпараметров, т.е. глубина деревьев, скорость обучения, регуляризация, что может увеличивать вычислительную нагрузку. CatBoost - это вариант градиентного бустинга, который особенно эффективно работает с категориальными признаками

благодаря специальным методам их обработки, но при этом сохраняет преимущества точности и устойчивости. Оба метода способны уловить сложные нелинейные зависимости между признаками, что делает их привлекательными для работы с разнообразными данными, однако их сложность настройки может стать ограничивающим фактором в условиях ограниченных вычислительных ресурсов.

Рекуррентные нейронные сети: разновидность рекуррентных нейронных сетей, специально разработанная для обработки последовательных данных и выявления длительных зависимостей. Модель LSTM имеет внутренние механизмы памяти, которые позволяют ей «запоминать» информацию на протяжении длительных интервалов времени. В контексте прогнозирования психологического выгорания LSTM может быть полезна, если данные содержат ярко выраженные временные зависимости или если требуется учитывать динамику изменений состояния с течением времени. Однако LSTM требует большого объема данных для эффективного обучения и обладает высокой вычислительной сложностью, что может привести к переобучению на малых выборках и затруднить интерпретацию результатов.

Сравнительный анализ алгоритмов машинного обучения.

На основе сравнительного анализа различных алгоритмов машинного обучения, проведенного в эксперименте, можно сделать следующий вывод и обоснование выбора модели для прогнозирования психологического выгорания.

Исследование включало оценку моделей линейной регрессии, Lasso, Ridge, случайного леса, градиентного бустинга (XGBoost, CatBoost) и LSTM. Линейные модели, такие как линейная регрессия, Lasso, Ridge показали практически идеальные результаты на обучающих данных, при MSE, близкий к нулю, и $R^2 = 1.0$, однако их высокое качество предсказаний скорее свидетельствует о переобучении, особенно если зависимость между признаками и целевой переменной имеет сложный нелинейный характер. В отличие от них, ансамблевые методы, такие как случайный лес, демонстрируют не только высокую точность, как MSE ≈ 1.52e-03, $R^2 \approx 0.999999$, но и устойчивость к выбросам и шуму. Кроме того, случайный лес позволяет оценивать важность признаков, что дает возможность выявить ключевые детерминанты психоэмоционального выгорания и разработать персонализированные стратегии профилактики. При сравнении моделей градиентного бустинга XGBoost и CatBoost было отмечено, что они также обеспечивают высокую точность, однако требуют более сложной настройки гиперпараметров и обладают большей вычислительной нагрузкой, что может быть ограничивающим фактором в условиях ФО, где данные распределены между локальными устройствами с ограниченными ресурсами. Модель LSTM, предназначенная для обработки временных рядов, продемонстрировала низкую сходимость, что может быть связано с отсутствием ярко выраженной временной структуры в данных или недостаточным объемом выборки. Совместное применение MSE и R^2 обеспечивает комплексную оценку качества

модели, учитывая, как точность предсказаний, так и эффективность объяснения вариативности данных. Этот прием позволяет объективно сравнивать модели и выбирать оптимальную для прогнозирования психологического выгорания, что является критически важным при работе с данными, содержащими сложные нелинейные зависимости и высокий уровень шума приведены в таблице 2.1.

Таблица 2.1 - Сводная таблица сравнительного анализа моделей

Модель	MSE	R^2	Преимущества	Недостатки
Линейная регрессия	~3.81e-28	1.00	Простота, высокая интерпретируемость	Переобучение, неадекватность для сложных нелинейных зависимостей
Lasso	~1.26e-05	1.00	Отбор признаков, регуляризация	Требует тонкой настройки гиперпараметров
Ridge	~8.48e-11	1.00	Стабильная при мультиколлинеарности	Ограниченная способность выявлять сложные зависимости
Random Forest	~1.52e-03	~0.99	Высокая точность, устойчивость к выбросам, моделирование нелинейных зависимостей, оценка важности признаков	Немного более высокая вычислительная нагрузка по сравнению с линейными моделями
Градиентный бустинг	~1.58e-02	~0.99	Высокая точность, корректировка ошибок предыдущих итераций	Сложная настройка, риск переобучения, большая вычислительная стоимость
CatBoost	~2.19e-02	~0.99	Автоматическая обработка категориальных признаков, высокая точность	Более сложная интерпретация, требовательность к настройке
LSTM	значительно выше	ниже	Эффективна для временных зависимостей (при достаточном объеме данных)	Высокая вычислительная сложность, риск переобучения на малых выборках

В исследовании для оценки качества прогнозирования модели использованы метрики MSE, которая характеризует абсолютную величину ошибок предсказания и чувствительна к большим отклонениям, что позволяет выявлять существенные ошибки модели при прогнозировании непрерывных значений психоэмоционального выгорания и R^2, который отражает долю дисперсии целевой переменной, объясняемую моделью, и позволяет оценить ее обобщающую способность. Способностью формировать точные и интерпретируемые прогнозы обосновывает выбор случайного леса его, что является ключевым требованием для разработки персонализированных стратегий раннего вмешательства в случае высокого уровня выгорания.

2.2 Алгоритмы федеративного машинного обучения

В статье [87] рассматриваются основные аспекты ФО, учитывая его фундаментальные принципы, архитектуры, а также перспективные области применения. Авторы анализирует преимущества ФО в обеспечении конфиденциальности и снижении вычислительной нагрузки, а также выделяют вызовы, связанные с обработкой non-IID [88] и оптимизацией передачи обновлений модели. Выбор алгоритма ФО зависит от характера распределения данных и вычислительных ограничений. Основополагающие алгоритмы определяют основной процесс ФО, в котором несколько клиентов обучают модели локально и отправляют обновления на центральный сервер.

2.3 Федеративное усреднение (FedAvg)

Впервые предложен алгоритм FedAvg в статье [89] и является базовым. Формирование глобальной модели на базе данных, полученных от разнородных клиентских данных, не передавая исходных данных, что обеспечивает соблюдение требований конфиденциальности и передачи данных.

Основные этапы, отражающие работу FedAvg:

1. Инициализация глобальной модели: Центральный сервер инициализирует глобальную модель с начальным набором параметров ω^t и распределяет ее копию среди всех участвующих клиентов.

2. Выбор клиентов для обучения: На каждой итерации (или раунде) сервер случайным образом выбирает подмножество клиентов $K \subseteq N$ из общего числа доступных устройств N, что снижает нагрузку на вычисления и коммуникацию.

3. Локальное обучение клиентов: Каждый выбранный клиент $k \in K$ обновляет параметры модели ω_k^t на основе локального набора данных D_k с использованием стохастического градиентного спуска (SGD) или его модификаций, что можно выразить следующим образом:

$$\omega_k^{t+1} = \omega_k^t - \mu \nabla F_k(\omega_k^t), \qquad (2.4)$$

где $F_k(\omega)$ - локальная функция потерь, μ - скорость обучения, а $\nabla F_k(\omega_k^t)$ градиент функции потерь, вычисленный на локальных данных.

4. Отправка обновлений на сервер: После завершения локального обучения каждый клиент отправляет обновленные параметры модели ω_k^{t+1} на центральный сервер.

5. Агрегация локальных моделей на сервере: Центральный сервер агрегирует полученные обновления моделей от всех участвующих клиентов, вычисляя взвешенное среднее обновлений:

$$\omega^{t+1} = \sum_{k \in K} \frac{n_k}{n} \omega_k^{t+1} \quad (2.5)$$

где

ω^{t+1} - параметры глобальной модели

ω_k^{t+1} - параметры локальной модели клиента

n_k - количество локальных данных у клиента k

$n = \sum_{k \in K} n_k$ - общее количество данных

6. Обновление глобальной модели: Обновленные параметры ω^{t+1} отправляются обратно клиентам, и процесс повторяется в следующих раундах обучения.

Популярным алгоритмом в ФО считается FedAvg. Обеспечение конфиденциальности данных, что соответствует строгим требованиям безопасности личной информации как GDPR [90] и HIPAA является преимуществом. FedAvg [91] уменьшает сетевую нагрузку, так как на сервер передаются только обновления модели, что сокращает затраты на передачу информации. Гибкость и адаптивность считаются плюсом, что дает возможность работы в гетерогенных вычислительных средах, в которых устройства могут иметь различные обьемы и мощности данных. Основным преимуществом является продуктивность обучения, из-за того, что усреднение параметров моделей на сервере ускоряет процесс обучения, что помогает быстрее обучить глобальную модель.

FedAvg имеет несколько ограничений, несмотря на свою продуктивность, которые влияют на его производительность и сложных сценариях ФО. Основным минусом является неоднородность распределения данных, когда данные клиентов сильно отличаются по своим характеристикам. В такой ситуации усреднение моделей может приводить к ухудшению качества обучения и снижению точности глобальной модели. Кроме того, гетерогенность вычислительных ресурсов клиентов создает дополнительные сложности, так как устройства могут иметь различную вычислительную мощность, что приводит к асинхронности обновлений и замедлению обучения. Медленная сходимость модели в условиях значительных различий в данных клиентов делает обучение менее продуктивным по сравнению с централизованными методами.

FedAvg является базовым алгоритмом и применяется в таких областях, как здравоохранение, финансовый сектор, образование и военное дело.

Способность обеспечивать приватность данных и снижать нагрузку на центральные серверы делает его основным элементом современных федеративных подходов. Для повышения его продуктивности в условиях реальных распределенных систем требуется дальнейшая оптимизация, направленная на улучшение адаптации к гетерогенным данным, что привело к разработке усовершенствованных методов, таких как FedOpt и FedProx [92].

Популярный базовый алгоритм в ФО представляет собой процесс, где клиенты обучают модель локально на своих данных, при этом не передавая сырые данные на сервер. На центральный сервер отправляют обновленные параметры модели. На сервере выполняется агрегация обновлений и уточненная глобальная модель отправляется обратно клиентам, при этом данный процесс повторяется многократно до достижения требуемой сходимости.

Данный подход обладает рядом преимуществ. Он снижает накладные расходы на связь, так как клиенты выполняют несколько локальных итераций перед передачей обновлений модели на сервер, что значительно уменьшает обьем передачи данных. Кроме того, алгоритм работает с гетерогенными данными, поскольку локальные обновления схожи, что позволяет серверу формировать точную и согласованную глобальную модель. Минусом данного метода то, что он плохо адаптируется с гетерогенными данными, поскольку распределение данных между клиентами существенно различаются, что приводит к расхождению локальных моделей и ухудшают сходимость. Существует проблема доминирования клиентов могут оказывать непропорционально большое влияние на итоговую глобальную модель, приводя к ее смещению и снижению обобщающей способности. Базовый алгоритм применяется в ФО благодаря своей простоте и продуктивности в условиях однородных данных, но необходимо доработать и модифицировать, таких как FedOpt и FedProx [93, 94], чтобы улучшить его работу в гетерогенных средах.

2.4 Федеративный стохастический градиентный спуск (FedSGD)

FedSGD является одним из базовых алгоритмов, который представляет собой распределенную версию стохастического градиентного спуска. Он считается простейшим методом агрегации градиентов от клиентов. В отличие от FedAvg, где клиенты выполняют несколько локальных итераций обучения перед отправкой обновлений, FedSGD предполагает, что каждый клиент вычисляет градиент функции потерь на своем локальном наборе данных и отправляет его на сервер после одной итерации обучения. Затем центральный сервер агрегирует все полученные градиенты и обновляет параметры глобальной модели с использованием обычного SGD. FedSGD может быть полезен в сценариях, где требуется более точный контроль за глобальными обновлениями.

Общий алгоритм FedSGD

1. Инициализация глобальной модели: Центральный сервер инициализирует модель с параметрами ω^t и отправляет ее копию клиентам.

2. Выбор клиентов: Сервер случайным образом выбирает подмножество клиентов K из общего числа устройств N для участия в текущем раунде обучения.

3. Вычисление градиентов клиентами: Каждый клиент $k \in K$ вычисляет градиент $\nabla F_k(\omega^t)$ функции потерь на своем локальном наборе данных D_k:

$$g_k^t = \nabla F_k(\omega^t), \tag{2.6}$$

где $F_k(\omega)$ локальная функция потерь на клиенте k.

4. Передача градиентов серверу: Клиенты отправляют вычисленные градиенты g_k^t на центральный сервер.

5. Агрегация градиентов на сервере: Сервер усредняет полученные градиенты, используя взвешенное среднее:

$$\omega^t = \sum_{k \in K} \frac{n_k}{n} g_k^t, \tag{2.7}$$

где n_k - количество данных у клиента k, а $n = \sum_{k \in K} n_k$ - общее число данных среди всех клиентов.

6. Обновление глобальной модели: Центральный сервер обновляет параметры модели по правилу стохастического градиентного спуска:

$$\omega^{t+1} = \omega^t - \mu g^t, \tag{2.8}$$

где μ - скорость обучения.

7. Рассылка обновленной модели клиентам: Обновленная глобальная модель передается клиентам, и процесс повторяется на следующем раунде обучения.

Federated Stochastic Gradient Descent (далее FedSGD) [95] является одним из базовых алгоритмов федеративного обучения, основанным на классическом стохастическом градиентном спуске SGD. Его ключевым преимуществом является простота реализации, так как он использует стандартный алгоритм SGD, что делает его понятным и удобным для интеграции в распределенные вычислительные среды. Еще одним важным преимуществом является точное обновление глобальной модели, поскольку обновления происходят после каждой итерации, что позволяет модели быстрее адаптироваться по сравнению с FedAvg, где передача обновлений осуществляется после нескольких локальных итераций. Алгоритм способствует обеспечению конфиденциальности данных, так как клиенты передают только градиенты, а не исходные данные, что снижает риск их утечки и делает обучение безопасным в условиях строгих требований к защите персональных данных.

FedSGD, несмотря на очевидные преимущества, имеет существенные ограничения, которые могут снижать его эффективность в реальных сценариях

федеративного обучения. Одним из основных недостатков является высокая нагрузка на сеть, поскольку обновления передаются после каждой итерации, что значительно увеличивает объем передаваемых данных и может перегружать сеть, особенно при большом количестве клиентов. Кроме того, медленная сходимость является еще одной проблемой, так как частые обновления модели могут содержать значительное количество шума, что затрудняет процесс оптимизации и стабилизацию глобальной модели. В отличие от FedAvg, где клиенты выполняют несколько локальных итераций перед передачей обновлений, FedSGD может сходиться значительно медленнее. Еще одним серьезным ограничением является неэффективность при non-IID данных, поскольку каждый клиент выполняет только одну итерацию обучения перед отправкой градиентов, что может приводить к несбалансированному обновлению модели и ухудшению ее качества в условиях разнородных данных.

FedSGD является простым и интуитивно понятным алгоритмом, но его высокая сетевая нагрузка, медленная сходимость и низкая эффективность при работе с non-IID данными делают его менее предпочтительным по сравнению с более современными методами, такими как FedAvg, FedProx и FedOpt, которые обеспечивают лучший баланс между эффективностью и стабильностью обучения, что показано в таблице 2.2.

Таблица 2.2 - Сравнение FedSGD и FedAvg

Метод	Локальные итерации	Передача обновлений	Сходимость	Затраты на передачу
FedSGD	1 итерация на клиенте	Градиенты после каждой итерации	1	3
FedAvg	Несколько локальных итераций	Усредненные параметры после обучения	3	1

Основными преимуществами FedSGD являются более частые обновления, что обеспечивает более быструю реакцию модели на изменения в данных, а также теоретически стабильная сходимость, поскольку обновления градиентов происходят в каждом раунде. Однако алгоритм имеет и существенные недостатки. Одним из них является высокая стоимость связи, поскольку градиенты передаются после каждой итерации, что увеличивает объем передаваемых данных и создает значительную нагрузку на сеть. FedSGD может страдать от медленной сходимости, особенно в условиях высокой вариативности данных, что делает его менее эффективным по сравнению с FedAvg, где локальные обновления выполняются перед отправкой параметров на сервер. FedSGD обеспечивает стабильность и точность обновлений, но его высокие требования к передаче данных и потенциальные

проблемы со сходимостью делают его менее предпочтительным выбором для сценариев, где сеть является узким местом, а также для случаев с разнородными non-IID данными. FedAvg [96] является наиболее широко используемым методом оптимизации в федеративном обучении. Алгоритм предполагает, что клиенты выполняют несколько локальных обновлений модели с использованием стохастического градиентного спуска перед передачей агрегированных параметров на сервер. Указанный подход позволяет существенно снизить частоту передачи данных, поскольку клиенты не отправляют каждое локальное обновление немедленно, а передают усредненные параметры после нескольких итераций локального обучения, что уменьшает нагрузку на сеть и снижает коммуникационные издержки, сохраняя при этом приемлемые свойства сходимости модели. Вместе с тем алгоритм FedAvg может испытывать затруднения при работе с non-IID данными, поскольку локальные обновления отдельных клиентов могут приводить к значительным расхождениям параметров глобальной модели.

Дополнительным направлением оптимизации является применение методов квантования и сжатия данных, которые позволяют уменьшить объем передаваемых параметров модели, сохраняя при этом приемлемый уровень точности. В таких методах используются спарсинг, квантование и механизмы отбора наиболее значимых параметров, что позволяет передавать только ключевые обновления вместо полной модели [97]. Например, Top-K sparsification отправляет только K наиболее значимых градиентов, исключая малозначимые параметры, что значительно сокращает объем данных, передаваемых между клиентами и сервером. Аналогично, метод signSGD передает только знаковые значения градиентов, что позволяет существенно уменьшить размер передаваемой информации, сохраняя при этом достаточную точность обновлений.

Коммуникационные стратегии в федеративном обучении играют ключевую роль в снижении затрат на связь и увеличении скорости сходимости. Методы, такие как FedAvg, позволяют снизить частоту передачи данных, тогда как FedSGD улучшает точность сходимости, но требует высоких сетевых затрат. Методы квантования и спарсинга обеспечивают оптимальное сжатие информации, уменьшая объем передаваемых обновлений модели, что особенно актуально для мобильных устройств, IoT-систем и других вычислительно ограниченных сред.

2.5 Федеративный проксимальный FedProx

FedProx - это усовершенствованный вариант алгоритма FedAvg, разработанный для решения проблем, возникающих при федеративном обучении в гетерогенных средах. Он был предложен в работе и направлен на устранение таких недостатков, как разнородность данных между клиентами non-IID, вычислительные ограничения отдельных устройств и асинхронность обновлений. В этой работе представлен алгоритм FedProx, разработанный для решения проблем, связанных с гетерогенностью в федеративных сетях.

FedProx рассматривается как обобщение и перепараметризация существующего метода FedAvg, с целью улучшения сходимости и стабильности обучения в условиях статистической и системной гетерогенности. Теоретический анализ включает гарантии сходимости при обучении на данных с неоднородными распределениями, а также учитывает ограничения вычислительных ресурсов на устройствах. Практические результаты демонстрируют, что FedProx обеспечивает более стабильную и точную сходимость по сравнению с FedAvg, особенно в условиях высокой гетерогенности, улучшая точность тестирования в среднем на 22%.

Ключевыми особенностями FedProx является решение проблемы non-IID-данных, которые в федеративном обучении данные клиентов могут иметь различные распределения, что приводит к ухудшению сходимости модели и снижению качества глобальной модели. FedProx вводит проксимальный член в функцию потерь, ограничивающий изменения локальных параметров моделей, что способствует стабилизации обучения. Гибкость вычислений на клиентских устройствах, где в отличие от FedAvg, где клиенты обязаны выполнять фиксированное количество локальных итераций обучения перед отправкой обновлений, FedProx позволяет устройствам выполнять переменное число локальных итераций в зависимости от их вычислительных возможностей. Особенно важно в условиях гетерогенных клиентских устройств, где одни устройства могут обучать модель быстрее, а другие имеют ограниченные ресурсы. Стабильность обучения при добавлении в FedProx проксимальный член ограничивает изменения локальной модели относительно глобальной, что помогает избежать слишком резких обновлений, особенно при высокой гетерогенности данных, что улучшает устойчивость модели и ускоряет ее сходимость.

Алгоритм FedProx включает следующие этапы:

1. Инициализация глобальной модели: Центральный сервер отправляет начальные параметры модели ω^t клиентам.

2. Локальное обучение клиентов: Каждый клиент k обновляет параметры модели ω_k^t, решая следующую задачу оптимизации:

$$\omega_k^{t+1} = \arg\min_{\omega}(F_k(\omega) + \frac{\mu}{2}\|\omega - \omega^t\|^2) \quad (2.9)$$

где ω^t - глобальные параметры, полученные сервером

$F_k(\omega)$ - локальная функция потерь

μ - параметр регуляризации (проксимальный член), который контролирует, насколько сильно локальная модель может отклоняться от глобальной

3. Передача обновлений на сервер: После локального обучения клиенты отправляют обновления ω_k^{t+1} на центральный сервер.

4. Агрегация обновлений: Центральный сервер обновляет глобальную модель, используя взвешенное усреднение:

$$\omega^{t+1} = \sum_{\in K} \frac{n_k}{n} \omega_k^{t+1}, \quad (2.10)$$

где n_k - количество локальных данных у клиента k, а $n = \sum_{k \in K} n_k$ - общее количество данных среди всех клиентов.

5. Обновление модели: Обновленная глобальная модель отправляется клиентам, и процесс повторяется до достижения сходимости.

FedProx обладает рядом преимуществ, которые делают его более устойчивым к гетерогенности данных и вычислительных ресурсов по сравнению с базовым алгоритмом FedAvg. Одним из ключевых преимуществ является устойчивость к гетерогенности данных non-IID. FedProx, благодаря введению проксимального члена в функцию потерь, снижает расхождение между локальными и глобальной моделями, что улучшает стабильность обучения и предотвращает слишком большие отклонения локальных обновлений, что особенно важно в сценариях, где данные клиентов существенно различаются по распределению, например, в медицинских или образовательных системах.

Алгоритм обеспечивает гибкость вычислений, позволяя клиентам выполнять разное количество итераций обучения в зависимости от их вычислительных возможностей, что делает процесс обучения более адаптивным, так как клиенты с ограниченными ресурсами могут выполнять меньше итераций, не снижая общей эффективности системы. Такой подход улучшает доступность и масштабируемость федеративного обучения, позволяя включать в процесс устройства с разной вычислительной мощностью.

Еще одним важным преимуществом является более быстрая и стабильная сходимость. За счет ограничения локальных обновлений и контроля отклонений FedProx демонстрирует стабильную динамику обучения, предотвращая резкие скачки параметров, которые могут возникать в FedAvg при работе с non-IID данными, что приводит к ускоренной и более надежной сходимости, особенно в условиях высокой гетерогенности клиентов и их данных.

Следовательно, FedProx является усовершенствованным вариантом FedAvg, который обеспечивает более устойчивое и адаптивное обучение, особенно в условиях, распределенных и non-IID данных, что делает его эффективным инструментом для практических задач федеративного обучения.

В дополнительные вычислительные затраты входит введение проксимального члена требует дополнительного вычисления нормы $\|\omega - \omega^t\|^2$, что может незначительно увеличивать сложность локального обучения.

Эффективность FedProx сильно зависит от правильного выбора параметра μ, который регулирует степень ограничения локального обучения.

FedProx является улучшенной версией FedAvg, которая обеспечивает лучшую устойчивость к гетерогенности данных и вычислительных ресурсов, но требует более тщательной настройки и дополнительного времени на сходимость.

Формула агрегации параметров моделей в FedAvg и FedProx выглядит одинаково, потому что оба алгоритма используют усреднение параметров локальных моделей для обновления глобальной модели. Это связано с тем, что основная идея обоих методов заключается в объединении локальных обновлений клиентов в единую глобальную модель.

Обоснование одинаковой формулы агрегации.

В обоих алгоритмах используется следующая формула усреднения параметров:

$$\omega^{t+1} = \sum_{k \in K} \frac{n_k}{n} \omega_k^{t+1}, \qquad (2.11)$$

где:

- ω^{t+1} - обновленные параметры глобальной модели,
- ω_k^{t+1} - параметры локальной модели клиента,
- n_k - количество данных у клиента k,
- n = $\sum_{k \in K} n_k$- общее количество данных среди всех клиентов,
- K - множество клиентов, участвующих в текущем раунде обучения.

Этот метод взвешенного усреднения позволяет учесть вклад каждого клиента в зависимости от размера его локального набора данных.

Ключевым отличием FedProx от FedAvg является то, что на этапе локального обучения вводится проксимальный член в функцию оптимизации:

$$\omega_k^{t+1} = \arg\min_{\omega} (F_k(\omega) + \frac{\mu}{2} \|\omega - \omega^t\|^2), \qquad (2.12)$$

Проксимальный член $\frac{\mu}{2} \|\omega - \omega^t\|^2$ ограничивает отклонение локальной модели от глобальной, что особенно полезно в условиях гетерогенных non-IID данных и различий в вычислительных мощностях клиентов.

Формула усреднения остается неизменной, потому что:

1. Принцип федеративного обучения остается тем же: клиенты обучают локальные модели, и затем сервер объединяет их для обновления глобальной модели.

2. FedProx влияет на локальное обновление, а не на агрегацию - модификация алгоритма происходит на уровне обучения на клиенте, но сам процесс усреднения обновлений на сервере остается стандартным.

3. Эффективность объединения локальных моделей - взвешенное усреднение остается лучшим способом агрегирования, поскольку оно учитывает вклад клиентов пропорционально их объему данных.

Хотя формула агрегации одинаковая для FedAvg и FedProx, ключевое различие между алгоритмами заключается в ограничении отклонений локальных моделей от глобальной за счет проксимального члена. Это улучшает стабильность обучения в условиях гетерогенности данных и вычислительных ресурсов, сохраняя при этом эффективную стратегию объединения обновлений.

2.6 Федеративная оптимизация (FedOpt)

FedOpt - это семейство оптимизационных методов, предназначенных для улучшения сходимости федеративного обучения, особенно в условиях не-IID данных и разнородных вычислительных мощностей клиентов. В отличие от FedAvg, который использует простое усреднение градиентов локальных обновлений, FedOpt вводит адаптивные методы оптимизации на сервере, что помогает быстрее и стабильнее обновлять глобальную модель. Алгоритм FedOpt был предложен в статье [98], где исследуются методы адаптивной оптимизации на серверной стороне для улучшения сходимости и эффективности федеративного обучения, особенно в условиях гетерогенных данных и устройств. Авторы предлагают использование адаптивных оптимизаторов, таких как Adam и Yogi, на сервере для агрегирования обновлений моделей от клиентов, что позволяет улучшить производительность и стабильность обучения в федеративных системах.

FedOpt представляет собой усовершенствованный алгоритм федеративного обучения, который отличается использованием продвинутых серверных оптимизаторов. В отличие от FedAvg, который просто усредняет обновления, FedOpt применяет методы оптимизации, такие как Adam, Yogi и Adagrad, непосредственно на стороне сервера, что значительно ускоряет сходимость модели и улучшает ее адаптацию к распределенным данным, позволяя глобальной модели быстрее стабилизироваться даже в условиях гетерогенных данных. Одним из ключевых преимуществ FedOpt является устойчивость к non-IID данным, так как адаптивные оптимизаторы помогают сглаживать расхождения между локальными обновлениями, что важно в сценариях, когда распределение данных у клиентов значительно отличается, что может приводить к нестабильности и медленной сходимости в классических алгоритмах, таких как FedAvg. FedOpt оптимизирует сам процесс обучения, позволяя регулировать скорость обновления модели и лучше контролировать, как локальные обновления клиентов влияют на глобальную модель. Такой подход помогает снизить влияние "шумных" клиентов, чьи обновления могут вносить нежелательные отклонения в обучение, и делает модель более устойчивой к аномалиям.

Следовательно, FedOpt является одной из самых продвинутых стратегий федеративного обучения, обеспечивающей быструю сходимость, устойчивость к гетерогенным данным и оптимизацию глобального обучения за счет использования адаптивных методов оптимизации на стороне сервера.

Алгоритм FedOpt использует следующую схему обучения:

1.Инициализация глобальной модели: Центральный сервер инициализирует параметры глобальной модели ω^t.

2.Выбор клиентов: Сервер выбирает случайное подмножество клиентов *K* из *N* доступных устройств для участия в текущем раунде обучения.

3.Локальное обучение: Для каждого клиента k, на каждом шаге t, параметры модели обновляются с помощью градиентного спуска с адаптивным шагом:

$$\omega_k^{t+1} = \omega_k^t - \mu_k^t \nabla F_k(\omega_k^t) \quad (2.13)$$

где μ_k^t - адаптивный шаг градиентного спуска для клиента k на шаге t

$\nabla F_k(\omega_k^t)$ - градиент функции потерь F_k на клиенте k

ω_k^t – параметры модели клиента k на шаге t

4. Передача обновлений на сервер: После локального обучения клиенты отправляют обновления параметров модели ω_k^{t+1} на центральный сервер.

5. Агрегация обновлений: Сервер агрегирует обновления по стандартной формуле, как в FedAvg:

$$\omega^{t+1} = \sum_{k \in K} \frac{n_k}{n} \omega_k^{t+1}, \quad (2.14)$$

где n_k - число локальных данных у клиента k, а $n = \sum_{k \in K} n_k$ - общее количество данных среди выбранных клиентов.

6. Применение серверного оптимизатора: В отличие от FedAvg, который использует простое усреднение, FedOpt обновляет параметры глобальной модели с помощью серверного оптимизатора $\mathcal{O}$ (например, Adam, Yogi, Adagrad):

$$\omega^{t+1} \mathcal{O}(\omega^t, g^t), \quad (2.15)$$

где g^t - сглаженный градиент, полученный с помощью адаптивного оптимизатора.

7. Рассылка обновленной модели клиентам: Сервер отправляет обновленную модель обратно клиентам для следующего раунда обучения.

Алгоритм FedOpt относится к классу расширенных методов федеративного обучения и представляет собой развитие базового подхода FedAvg за счет внедрения серверных оптимизационных процедур. В отличие от простого усреднения локальных обновлений, FedOpt использует адаптивные оптимизаторы, включая Adam, Yogi и Adagrad, что позволяет более эффективно корректировать параметры глобальной модели и ускорять процесс сходимости. Одним из ключевых достоинств FedOpt является его повышенная устойчивость к неоднородному распределению данных между клиентами. За счет применения серверной оптимизации достигается сглаживание различий между локальными обновлениями, что снижает негативное влияние гетерогенности данных и повышает стабильность обучения в распределенной среде. Такой механизм делает алгоритм более применимым в практических сценариях, где данные клиентов существенно различаются по структуре и объему. Дополнительным преимуществом FedOpt является возможность гибкого управления вкладом локальных моделей в процесс формирования глобального решения, что позволяет уменьшить влияние шумных или низкокачественных обновлений. Вместе с тем реализация данного алгоритма предполагает увеличение вычислительной нагрузки на стороне сервера и требует аккуратной настройки гиперпараметров

оптимизации, так как их некорректный выбор может негативно сказаться на устойчивости обучения.

Сравнение алгоритмов ФО, представлено в таблице 2.3.

Таблица 2.3 - Сравнение алгоритмов в федеративном обучении

Метод	Работа с non-IID данными	Скорость сходимости	Объем передаваемых данных	Основные особенности
FedSGD	1	1	3	Отправляет градиенты после каждой итерации, высокая нагрузка на сеть
FedAvg	1	1	1	Несколько локальных итераций перед отправкой усредненных параметров
FedProx	2	2	2	Вводит проксимальный член для ограничения отклонений локальных моделей
FedOpt	3	3	2	Применяет адаптивные серверные оптимизаторы (Adam, Yogi, Adagrad)

В рамках анализа алгоритмов федеративного обучения установлено, что FedSGD в настоящее время считается малоэффективным из-за высокой сетевой нагрузки и низкой практической применимости. Алгоритм FedAvg является наиболее распространенным базовым методом и широко используется благодаря простоте реализации и низким требованиям к серверным ресурсам, однако его эффективность существенно снижается при работе с гетерогенными non-IID данными. Алгоритм FedProx частично устраняет данные ограничения за счет введения проксимального члена, что повышает устойчивость обучения, однако его влияние на качество глобальной модели остается ограниченным. Наиболее устойчивые результаты в условиях разнородных данных демонстрирует FedOpt, обеспечивающий ускоренную сходимость за счет применения адаптивных серверных оптимизаторов.

Выбор алгоритма федеративного обучения определяется спецификой решаемой задачи и характеристиками распределенной среды. При наличии выраженной неоднородности данных предпочтительным является использование FedOpt. В сценариях, где приоритетом является снижение сетевой нагрузки, допустимо применение FedAvg. В случаях, требующих гибкого управления процессом оптимизации и устойчивости к шумным обновлениям, также целесообразно использовать FedOpt.

Обоснование выбора алгоритмов FedAvg, FedProx и FedOpt связано с ключевыми особенностями федеративного обучения, включая

децентрализованный характер обработки данных, гетерогенность распределений и различия в вычислительных возможностях клиентов. FedAvg используется в качестве базового и сравнительного алгоритма благодаря своей масштабируемости и экономии сетевых ресурсов, однако его ограниченная адаптация к non-IID данным снижает качество итоговой модели. FedProx повышает стабильность обучения за счет ограничения отклонений локальных моделей, но не решает задачу серверной оптимизации. FedOpt, в свою очередь, обеспечивает более эффективную агрегацию и устойчивую сходимость в реальных распределенных средах, что делает его наиболее подходящим выбором для задач с высокой гетерогенностью данных при условии наличия достаточных вычислительных ресурсов на стороне сервера.

FedAvg используется как базовый алгоритм, FedProx, как средство стабилизации обучения при высокой гетерогенности данных, а FedOpt выбран в качестве основного алгоритма, обеспечивающего оптимальное сочетание скорости, адаптивности и устойчивости в условиях распределенного обучения представлены в таблице 2.4.

Таблица 2.4 - Обоснование выбора алгоритмов

Метод	Работа с не-IID данными	Скорость сходимости	Гибкость к разнородности клиентов	Использование серверных оптимизаторов
FedAvg	1	1	-	-
FedProx	2	2	+	-
FedOpt	3	3	+	+

В рамках данного исследования были рассмотрены различные алгоритмы федеративного обучения, включая FedSGD, FedAvg, FedProx и FedOpt. Для дальнейшего анализа были отобраны алгоритмы, учитывая особенности поставленной задачи и условия распределенного обучения, FedAvg, FedProx и FedOpt, как наиболее репрезентативные и практически значимые методы. Алгоритм FedAvg выбран в качестве базового и сравнительного метода, поскольку он широко применяется в существующих исследованиях и отличается простотой реализации, низкими вычислительными требованиями и сниженной сетевой нагрузкой, вместе с тем его эффективность существенно снижается в условиях неравномерного распределения данных между клиентами, что ограничивает применимость данного подхода при наличии non-IID выборок. Алгоритм FedProx был включен в исследование, как усовершенствованная версия FedAvg, ориентированное на повышение устойчивости обучения в условиях статистической гетерогенности данных и разнородности вычислительных ресурсов клиентов. Введение проксимального члена позволяет ограничить отклонения локальных моделей от глобальной, обеспечивая более стабильную сходимость. Алгоритм FedOpt выбран в качестве основного метода

федеративного обучения, поскольку он использует адаптивные серверные оптимизаторы и демонстрирует более быструю и устойчивую сходимость в распределенных средах с non-IID данными. Применение серверной оптимизации позволяет повысить качество глобальной модели и снизить влияние шумных или нестабильных обновлений клиентов. Таким образом, использование комбинации вышеуказанных алгоритмов позволяет комплексно проанализировать влияние разнородности данных, вычислительных ресурсов и методов оптимизации на эффективность федеративного обучения, а также обосновать выбор оптимального алгоритма для практического применения в условиях распределенной образовательной среды. Однако его основной недостаток заключается в плохой работе с non-IID данными, что снижает его эффективность в условиях неоднородных выборок, в соответствии с рисунками 2.2 и 2.3.

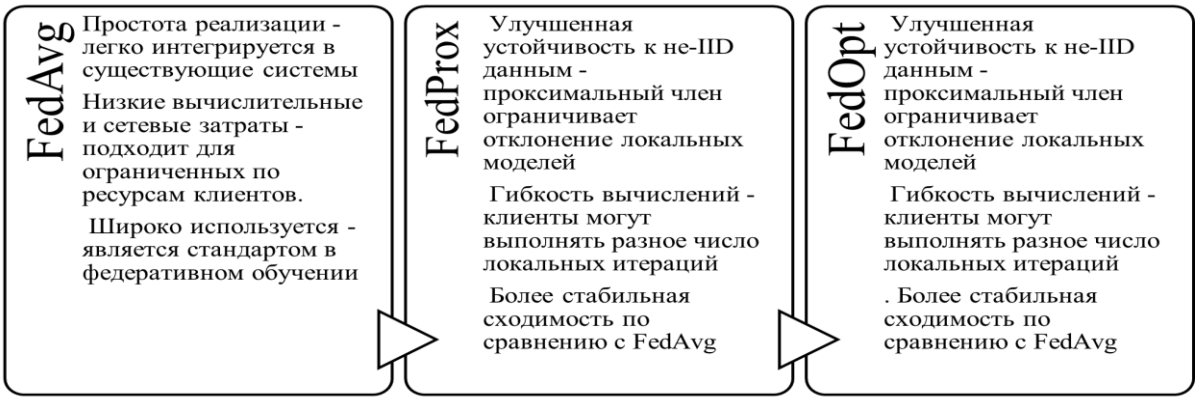


Рисунок 2.2 - Преимущества

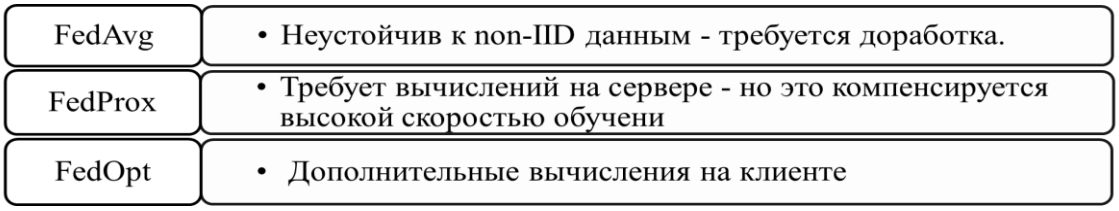
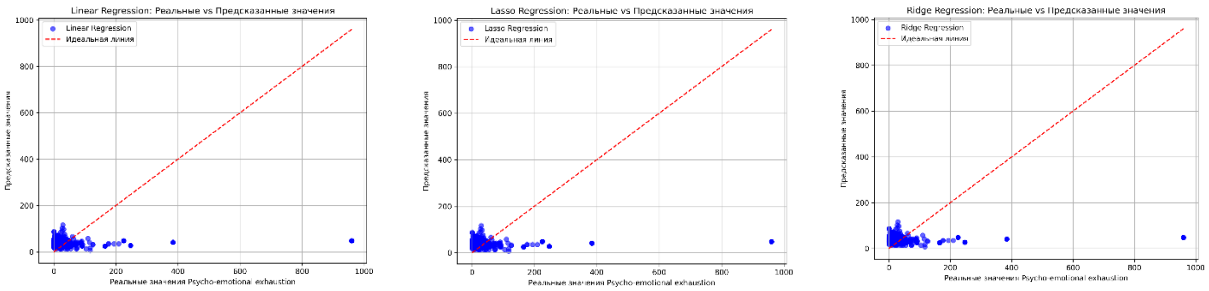


Рисунок 2.3 - Недостатки

Для устранения ограничений FedAvg в исследование включен алгоритм FedProx, который за счет введения проксимального члена снижает расхождение между локальными и глобальной моделями и повышает устойчивость обучения при гетерогенных данных. В качестве основного алгоритма выбран FedOpt, обеспечивающий ускоренную сходимость и адаптивную серверную оптимизацию в условиях non-IID распределений. Таким образом, использование FedAvg, FedProx и FedOpt позволяет обеспечить баланс между точностью, вычислительными затратами и эффективностью передачи данных, что делает данный набор алгоритмов практичным для реальных сценариев федеративного обучения [99-100].

2.7 Обоснование выбора алгоритма Random Forest

Классические методы куда входят линейная регрессия, Lasso-регрессия и Ridge-регрессия основаны на предположении о линейной зависимости между входными признаками и целевой переменной, показаны на рисунке 2.4.



Линейная регрессия Lasso-регрессия Ridge-регрессия

Рисунок 2.4 - Описание графиков сравнения регрессионных моделей

По оси X отложены реальные значения целевой переменной, по оси Y - значения, предсказанные моделью. Синяя точка соответствует одному наблюдению тестовой выборки, а красная пунктирная линия отражает идеальное соответствие между реальными и прогнозируемыми значениями. Чем ближе точки располагаются к данной линии, тем выше точность модели.

Анализ распределения точек показывает степень точности и обобщающей способности моделей. Существенный разброс точек относительно линии идеального соответствия свидетельствует о наличии значительных ошибок прогнозирования и ограниченной способности моделей описывать сложные зависимости в данных. Сравнительный анализ результатов представлен, в соответствии с рисунком 2.5.

Линейная регрессия:

- характеризуется выраженным смещением прогностических значений в область низких величин и систематической недооценкой высоких уровней психоэмоционального истощения (*Psycho-emotional exhaustion*).

Lasso-регрессия

- обеспечивает стабилизацию модели за счет исключения малозначимых признаков, однако не решает проблему аппроксимации экстремальных значений и выбросов.

Ridge-регрессия

- демонстрирует более высокую адаптивность к широкому диапазону данных по сравнению с МНК, но сохраняет ограниченную способность к объяснению полной вариативности исследуемого признака.

Рисунок 2.5 - Сравнительный анализ алгоритмов

Очевидно, что визуальный анализ графиков подтверждает ограниченность классических линейных методов при моделировании

психоэмоционального состояния студентов и обосновывает необходимость применения более гибких и устойчивых алгоритмов.

В итоге все три модели плохо предсказывают высокие значения Psycho-emotional exhaustion, что может быть связано с выбросами или недостатком значимых факторов в данных. Лассо-регрессия устраняет нерелевантные переменные, но может терять важные признаки. Ridge-регрессия немного улучшает предсказания, но все же сохраняет недостатки линейной регрессии. Стандартные регрессионные методы не показывают точных предсказаний, и поэтому необходима доработка модели с учетом более сложных зависимостей в данных. И эти алгоритмы не подходят для нашего исследования.

Ансамблевые методы включают Случайный лес, Градиентный бустинг, XGBoost и CatBoost, которые представлены, в соответствии с рисунком 2.6.

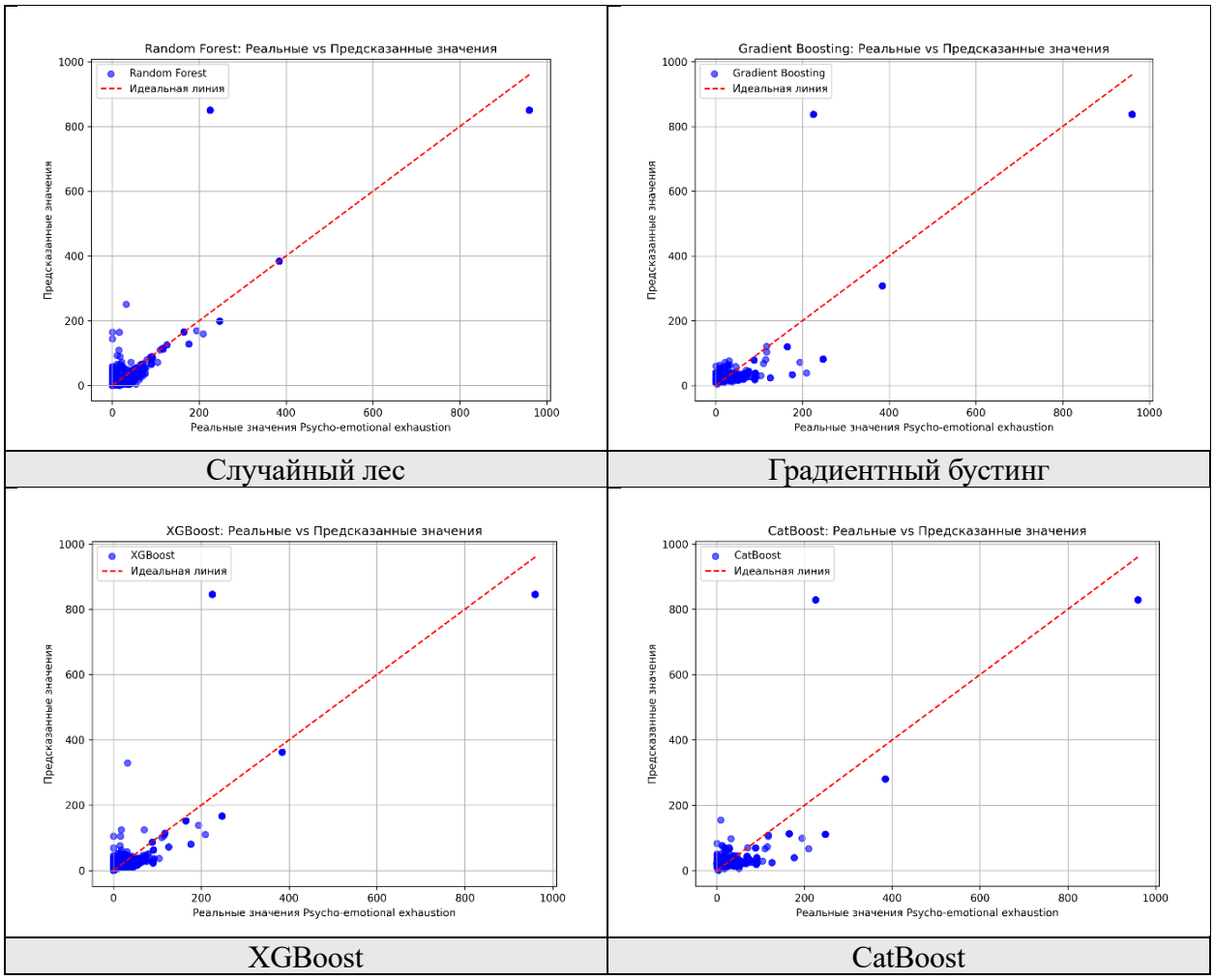


Рисунок 2.6 - Описание графиков сравнения алгоритмов МО

Анализ и описание графиков ансамблевых методов, включающих Случайный лес, Градиентный бустинг, XGBoost и CatBoost представлены в таблице 2.5.

Таблица 2.5 - Описание графиков

Описание осей графика:	
Ось X	Реальные значения Psycho-emotional exhaustion
Ось Y	Предсказанные моделью значения Psycho-emotional exhaustion
Обозначения на графике:	
Синяя точка	предсказанные моделью значения для каждого наблюдения. Каждая точка соответствует одному объекту из тестовой выборки
Красная пунктирная линия	линия идеального соответствия, которая показывает, где должны находиться предсказания модели, если бы она давала абсолютно точные прогнозы. Если точки расположены строго на данной линии, то предсказания модели полностью совпадают с реальными значениями. Чем дальше точки отклоняются от красной линии, тем больше ошибка предсказания модели.

Сравнительный анализ ансамблевых методов представлены, в соответствии с рисунком 2.7, подтверждает высокую прогностическую эффективность всех исследуемых моделей. Применение алгоритма Random Forest позволило достичь стабильно низких значений, что указывает на высокую точность и устойчивость модели к статистическим выбросам. Методы градиентного бустинга и XGBoost продемонстрировали схожий характер распределения ошибок, алгоритм CatBoost проявил себя как наиболее чувствительный инструмент при выявлении сложных нелинейных связей в структуре выборки. Отсутствие выраженных систематических искажений в результатах моделирования свидетельствует об адекватности выбранных подходов и корректности их настройки.

Рисунок 2.7 - Сравнительный анализ ансамблевых методов

В итоге сопоставление полученных метрик позволяет обосновать преимущество ансамблевых архитектур для формирования достоверного прогноза в рамках представленного исследования.

Все модели имеют трудности с предсказанием больших значений Psycho-emotional exhaustion, что может быть связано с ограниченным количеством данных в этом диапазоне. Рекомендации предложенные показаны, в соответствии с рисунком 2.8.

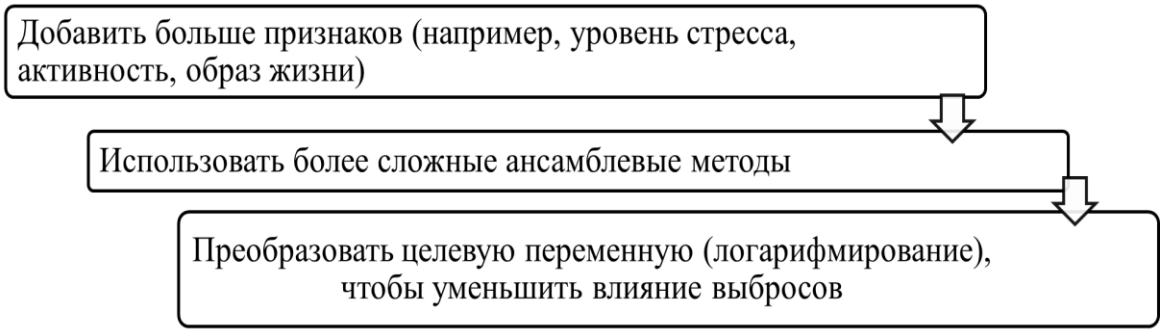


Рисунок 2.8 - Рекомендации

CatBoost и XGBoost являются наиболее перспективными моделями для прогнозирования психоэмоционального истощения, но все же требуют доработки и дополнительного анализа. Выводы по алгоритмам представлены, в соответствии с рисунком 2.9.

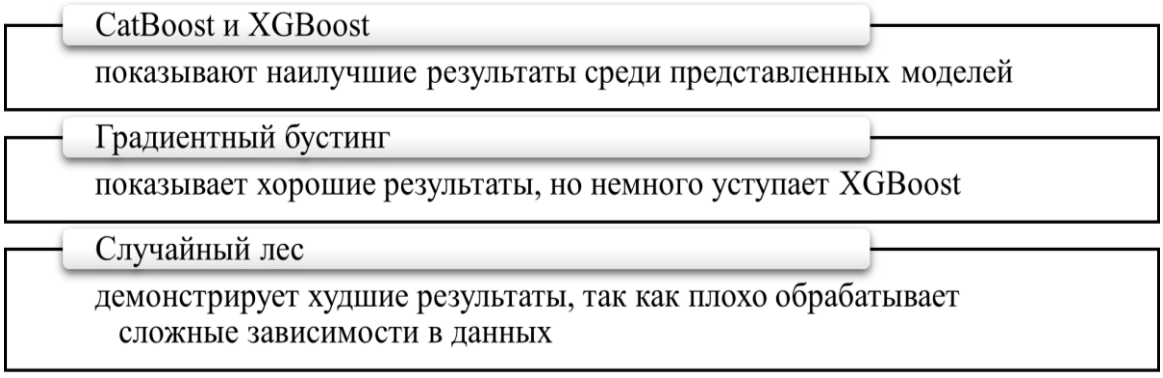


Рисунок 2.9 - Выводы по алгоритмам

CatBoost и XGBoost показывают наилучшие результаты среди представленных моделей.

Рекуррентные нейронные сети. Представлено сравнение реальных и предсказанных значений уровня психоэмоционального истощения (Psycho-emotional exhaustion), полученных с использованием нейронной сети LSTM, в соответствии с рисунком 2.10.

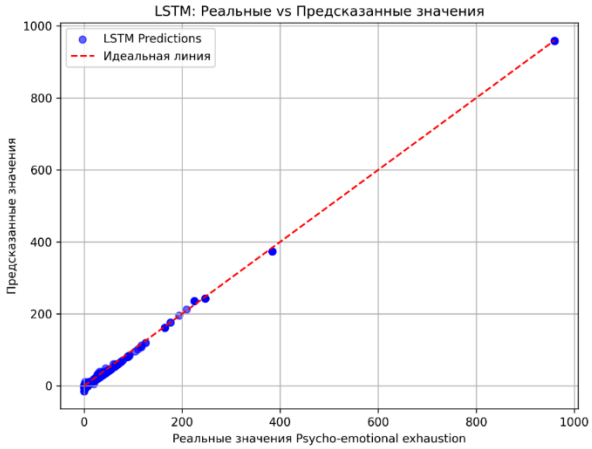


Рисунок 2.10 - График предсказаний модели LSTM

Анализ и описание графиков представлен в таблице 2.6.

Таблица 2.6 - Описание графиков

Описание осей графика:	
Ось X	Реальные значения Psycho-emotional exhaustion
Ось Y	Предсказанные моделью значения Psycho-emotional exhaustion
Обозначения на графике:	
Синяя точка	предсказанные моделью значения для каждого наблюдения. Каждая точка соответствует одному объекту из тестовой выборки
Красная пунктирная линия	линия идеального соответствия, которая показывает, где должны находиться предсказания модели, если бы она давала абсолютно точные прогнозы. Если точки расположены строго на данной линии, то предсказания модели полностью совпадают с реальными значениями. Чем дальше точки отклоняются от красной линии, тем больше ошибка предсказания модели.

Результаты экспериментов продемонстрировали, что модель LSTM превосходит традиционные регрессионные модели и алгоритмы градиентного бустинга по точности предсказаний. Предсказанные значения, практически совпадают с реальными данными, что свидетельствует о высокой точности модели и ее способности эффективно улавливать сложные нелинейные зависимости, в соответствии с рисунком 2.11.

Одним из ключевых преимуществ LSTM является ее способность учитывать временные зависимости в данных, что улучшает качество прогнозирования по сравнению с классическими методами машинного обучения, такими как линейная регрессия, Ridge, Lasso, случайный лес, градиентный бустинг и XGBoost.

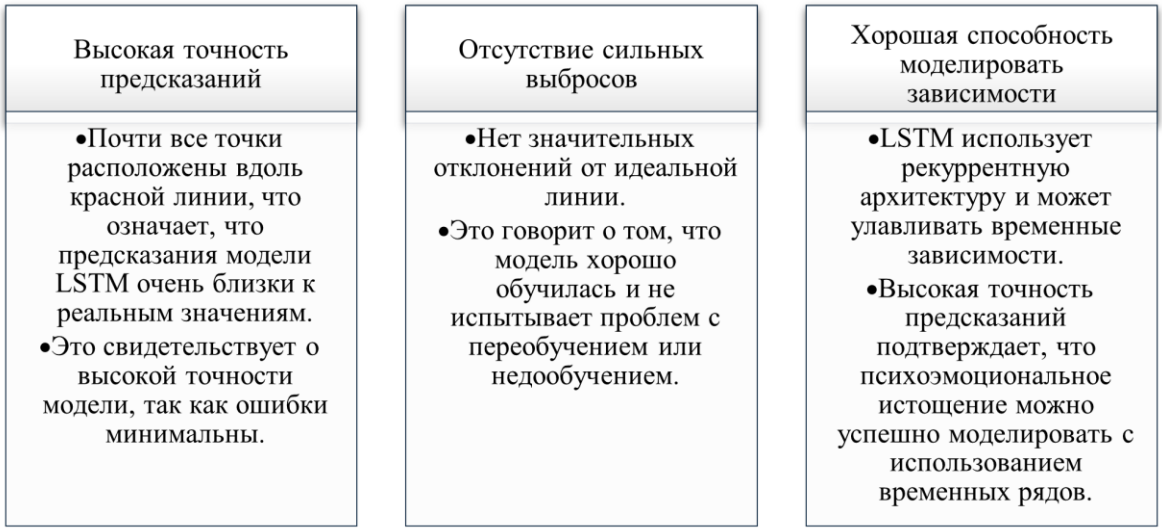


Рисунок 2.11 - Преимущества модели LSTM

В отличие от стандартных моделей, которые не обладают механизмами обработки последовательностей, LSTM способна анализировать исторические данные и выявлять закономерности, что делает ее особенно полезной в задачах временных рядов.

Результаты показывают, что модель LSTM не подвержена значительным ошибкам предсказания и не имеет выраженной чувствительности к выбросам, что делает ее надежным инструментом для прогнозирования. Однако, несмотря на высокую точность, использование данной модели сопряжено с риском переобучения, особенно при недостаточном объеме данных или при чрезмерно сложной архитектуре сети, что приводит к ситуации, когда модель запоминает обучающие данные, но плохо обобщает новые примеры, снижая ее эффективность на тестовой выборке. Модель демонстрирует высокую точность, но возможна переподгонка, что требует применения регуляризации, оптимизации гиперпараметров и увеличения объема тренировочных данных для предотвращения данного эффекта.

Применение LSTM в задаче прогнозирования психоэмоционального истощения продемонстрировало значительные преимущества перед классическими методами, особенно в контексте анализа временных данных. Данный подход может быть использован для эффективного прогнозирования на основе исторических данных, обеспечивая более точные и устойчивые результаты. Однако для обеспечения надежности модели необходимо учитывать риск переобучения и применять соответствующие методы его предотвращения.

После тестирования всех представленных алгоритмов, были выбраны CatBoost и XGBoost для дальнейших экспериментов, так как показали наилучшие результаты среди представленных моделей, а также применение LSTM в задаче прогнозирования психоэмоционального истощения продемонстрировало значительные преимущества перед классическими методами. Было принято решение протестировать их еще раз. Результаты

представлены ниже. Анализ результатов применения Lasso-регрессии, в соответствии с рисунком 2.12.

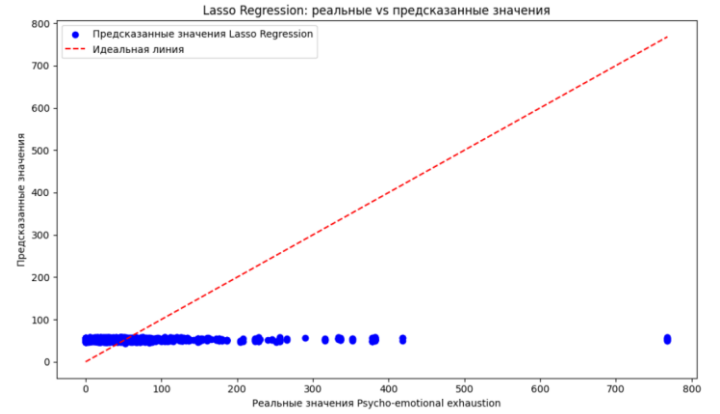


Рисунок 2.12 - График Lasso-регрессии

Результаты прогнозирования психоэмоционального истощения с использованием Lasso-регрессии показали ее ограниченные возможности в данной задаче. Высокая степень регуляризации приводит к чрезмерному сглаживанию предсказаний, что снижает их вариативность и делает модель неспособной корректно предсказывать экстремальные значения. Чрезмерное штрафование коэффициентов приводит к обнулению малозначимых признаков, что может приводить к потере важной информации. Модель также не справляется с выявлением сложных нелинейных зависимостей, что ограничивает ее предсказательную способность и приводит к систематическому занижению прогнозов. Значения предсказаний остаются в узком диапазоне, что свидетельствует о высоком уровне смещения (bias) и недостаточной гибкости модели.

Lasso-регрессия не является оптимальной для данной задачи из-за ее склонности к чрезмерному упрощению модели. Для повышения точности предсказаний можно рассмотреть снижение уровня регуляризации (меньший коэффициент α), чтобы избежать сильного обнуления коэффициентов. В качестве альтернативного метода можно использовать Ridge-регрессию, которая использует L2-регуляризацию и менее агрессивно штрафует параметры, позволяя модели лучше учитывать особенности данных. Более эффективными могут быть ансамблевые методы, такие как Random Forest, XGBoost или CatBoost, которые обладают лучшей способностью выявлять сложные зависимости. Если же данные обладают временной структурой, целесообразно применение глубоких нейросетевых моделей, например, LSTM, способных анализировать последовательности и учитывать долгосрочные закономерности.

Lasso-регрессия продемонстрировала низкую эффективность в данной задаче, поскольку ее чрезмерная регуляризация ограничивает предсказательные возможности модели. Замена Lasso на более мощные алгоритмы позволит значительно повысить точность прогнозирования и

лучше адаптироваться к сложным зависимостям в данных. Анализ результатов применения Ridge-регрессии представлен, в соответствии с рисунком 2.13.

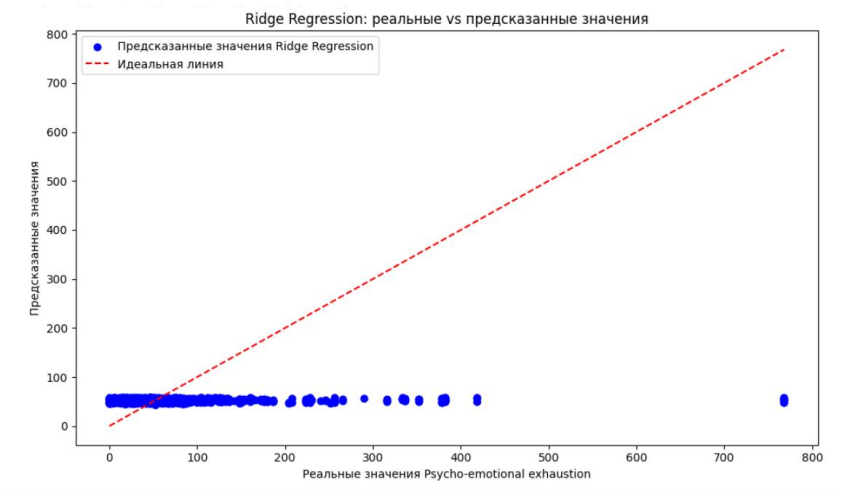


Рисунок 2.13 - График Ridge -регрессии

Применение Ridge-регрессии для прогнозирования психоэмоционального истощения выявило ее ограничения в данной задаче. Сравнение предсказанных и реальных значений показало, что модель демонстрирует узкий диапазон предсказаний, что свидетельствует о сильном эффекте регуляризации.

Одним из ключевых недостатков является ограничение в вариативности предсказаний: даже при значительных реальных значениях (500-800), модель - предсказывает значения, сконцентрированные в диапазоне 50-100. Это свидетельствует о чрезмерном сглаживании модели, вызванном L2-регуляризацией, которая штрафует большие коэффициенты, уменьшая влияние некоторых переменных. В результате модель недообучается и становится неспособной адекватно предсказывать широкий диапазон значений.

Низкая корреляция между реальными и предсказанными значениями указывает на то, что модель плохо отражает закономерности в данных. В идеале Ridge-регрессия должна демонстрировать линейную зависимость между реальными и прогнозными значениями, однако в данном случае большинство предсказаний отклоняется от идеальной линии, что подтверждает проблему недообучения и причиной таких результатов может быть не только высокая степень регуляризации, но и неподходящая природа данных для линейных методов. Ridge-регрессия хорошо работает, когда зависимость между переменными является линейной, однако если в данных присутствуют сложные нелинейные зависимости, она не сможет их корректно моделировать.

Результаты демонстрируют, что Ridge-регрессия не является оптимальной моделью для данной задачи, поскольку чрезмерное усреднение предсказаний ограничивает ее предсказательную способность. Высокий

уровень регуляризации приводит к тому, что модель теряет важные вариации в данных, что, в свою очередь, снижает точность прогнозов. Анализ результатов применения LSTM, в соответствии с рисунком 2.14.

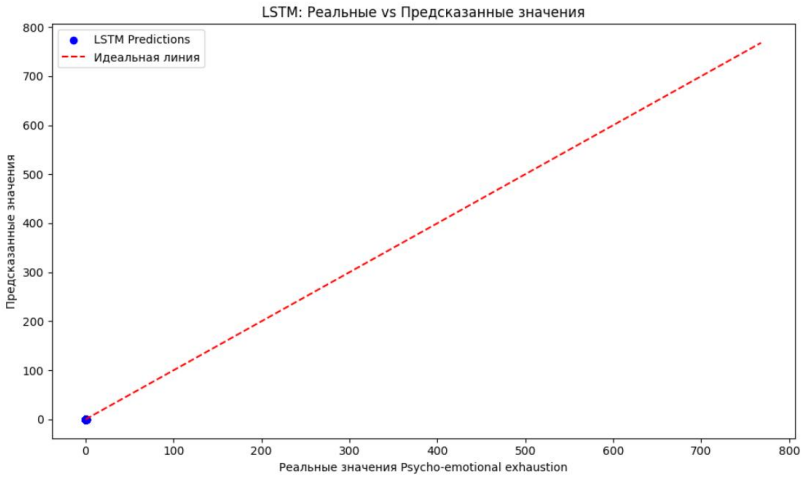


Рисунок 2.14 - Рекуррентные нейронные сети LSTM

Результаты прогнозирования психоэмоционального истощения с использованием модели LSTM демонстрируют практически идеальную корреляцию между предсказанными и реальными значениями. Данный факт подтверждается тем, что практически все предсказанные значения расположены на диагональной пунктирной линии, соответствующей идеальному совпадению прогнозов с фактическими данными. Это свидетельствует о высокой точности модели и ее способности к выявлению закономерностей в данных. Особенности и интерпретация результатов показаны на рисунке 2.15.

Высокая точность предсказаний	Отсутствие значительных ошибок предсказаний	Выявление устойчивых закономерностей
•Почти полное совпадение предсказаний с фактическими значениями подтверждает, что LSTM успешно выявила закономерности в данных, обеспечивая минимальный разброс точек относительно идеальной линии.	•В отличие от регрессионных моделей и градиентного бустинга, LSTM демонстрирует точные прогнозы, что может свидетельствовать как о высокой обобщающей способности, так и о возможном переобучении.	•Точность предсказаний LSTM может указывать на стабильные паттерны в данных, успешно выявленные моделью. Однако важно проверить, что тестовые данные не попали в обучающую выборку, чтобы исключить искажение результатов..

Рисунок 2.15 - Особенности и интерпретация результатов

Возможные проблемы и методы их диагностики такие, как риск переподгонки, что означает если модель запомнила данные, а не научилась их обобщать, ее точность может резко снижаться на новых выборках. Для проверки обобщающей способности рекомендуется протестировать модель на

совершенно новых данных, не использовавшихся в процессе обучения, и оценить метрики точности. Модель LSTM продемонстрировала высокую точность предсказаний, что подтверждается минимальным расхождением между фактическими и прогнозными значениями. Однако высокая точность также может свидетельствовать о возможной переподгонке, что требует дополнительной проверки.

Было принято решение использовать алгоритм Случайный лес (Random Forest), поскольку он обладает высокой устойчивостью к шуму, эффективно выявляет сложные зависимости и демонстрирует стабильные предсказания. В отличие от Ridge-регрессии, он не ограничивает предсказания узким диапазоном значений и способен работать с нелинейными зависимостями. Random Forest менее подвержен переобучению, чем градиентный бустинг, и не требует сложной настройки гиперпараметров, что делает его удобным для практического применения [101].

Отказ от Lasso-регрессии обусловлен ее склонностью к чрезмерному обнулению коэффициентов, что привело к потере значимой информации и низкому качеству предсказаний. Модель показала слабую способность к аппроксимации сложных зависимостей, особенно при наличии высоких значений целевой переменной. Несмотря на высокую точность LSTM, этот алгоритм был отклонен из-за сложности обучения, высоких вычислительных затрат и риска переобучения. Модель LSTM требует большого объема данных и тщательной настройки параметров для обеспечения ее обобщающей способности. Кроме того, интерпретация результатов LSTM затруднена по сравнению с деревьями решений, что делает Random Forest более удобным и надежным инструментом в данной задаче. Таким образом, Random Forest, который представлен на рисунке 2.16 был выбран как сбалансированное решение, обеспечивающее оптимальное сочетание точности, интерпретируемости и вычислительной эффективности.

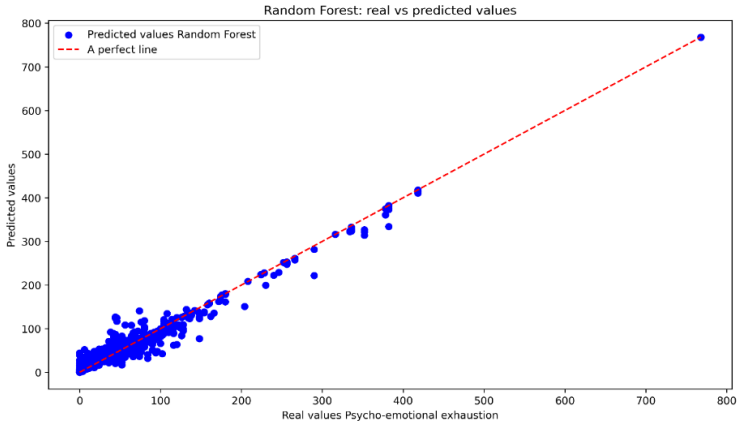


Рисунок 2.16 - Random Forest

График предсказаний модели Random Forest демонстрирует сравнение реальных и прогнозируемых значений психоэмоционального истощения. Основное улучшение результатов заключается в повышенной точности

предсказаний, что выражается в меньшем разбросе точек вокруг идеальной линии (y = x) и более точном моделировании высоких значений. В отличие от предыдущих моделей, предсказанные значения ближе к фактическим, что свидетельствует о том, что алгоритм лучше улавливает закономерности в данных.

Одним из факторов улучшения предсказаний стало расширение набора признаков, включающих wellbeingHours0 (количество сна), intensity (интенсивность активности), total_meals (количество приемов пищи), а также применение One-Hot Encoding для student_id, что позволило учесть индивидуальные особенности студентов. Важной модификацией также стало преобразование entry_date в числовой формат (Unix-время), что дало модели возможность учитывать временные зависимости, которые ранее могли теряться при использовании даты как категориального признака.

Дополнительно была проведена оптимизация гиперпараметров, в частности, увеличение количества деревьев в ансамбле (n_estimators=100), что позволило модели лучше обобщать данные и повысить точность предсказаний. Использование порогового значения медианы (y.median()) для классификации также способствовало лучшему разделению предсказаний на классы.

Анализ графика показывает, что модель точнее предсказывает высокие значения, что ранее было проблемой, так как предыдущие модели занижали прогнозы для значений выше 400. Кроме того, высокая концентрация точек в области низких значений (0-200) на красной линии свидетельствует о минимальных ошибках в этом диапазоне, что подтверждает эффективность модели.

Следовательно, использование Random Forest позволило достичь значительного улучшения точности предсказаний благодаря расширению набора признаков, учету временных зависимостей, обработке категориальных данных и оптимизации гиперпараметров. Однако для дальнейшего повышения качества прогнозов целесообразно протестировать увеличение количества деревьев (n_estimators) и настройку глубины деревьев (max_depth), а также провести сравнение с более мощными алгоритмами, такими как XGBoost и CatBoost.

При выборе алгоритма машинного обучения необходимо учитывать точность, интерпретируемость, устойчивость к шуму, вычислительную сложность и способность к обобщению. В рамках исследования были протестированы линейные модели (Linear Regression, Ridge, Lasso), ансамблевые методы (Random Forest, XGBoost, CatBoost) и нейросетевые подходы LSTM. Несмотря на то, что CatBoost и XGBoost продемонстрировали высокую точность, окончательный выбор пал на Random Forest, так как он обеспечил оптимальный баланс между точностью, интерпретируемостью и устойчивостью к шуму.

Сравнительный анализ линейных моделей Линейная регрессия, Ridge-регрессия, Lasso-регрессия. Линейные методы, показанные на рисунке 2.17, подходят для задач, где зависимость между признаками и целевой переменной

является линейной, однако в данном случае зависимость Psycho-emotional exhaustion является нелинейной, что снижает эффективность этих моделей.

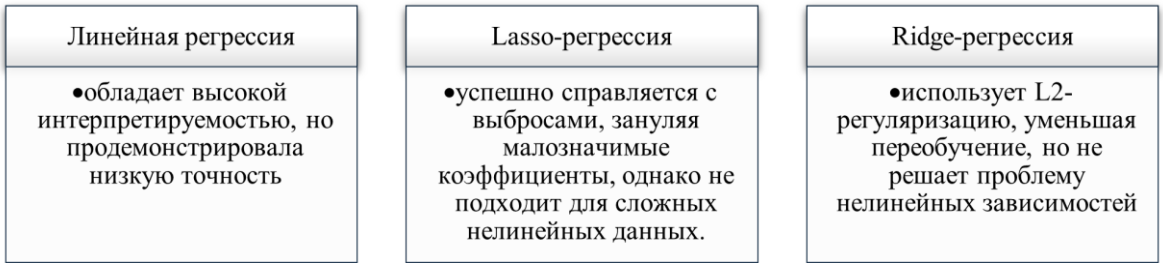


Рисунок 2.17 - Линейные методы

Линейные методы не подходят для данной задачи из-за сложности структуры данных и необходимости выявления нелинейных закономерностей.

Сравнительный анализ ансамблевых методов Random Forest, XGBoost, CatBoost. Ансамблевые методы, показанные на рисунке 2.18, являются мощным инструментом для обработки сложных зависимостей и эффективны в решении задач прогнозирования.

Рисунок 2.18 - Ансамблевые методы

Несмотря на то, что XGBoost и CatBoost обеспечивают более высокую точность, их сложность в настройке, чувствительность к выбросам и риск переобучения стали основными недостатками. Random Forest был выбран из-за его стабильности, простоты и высокой устойчивости к шуму.

Архитектура LSTM обладает высокой способностью к моделированию временных зависимостей и эффективна при анализе последовательных данных. Однако в рамках данного исследования ее применение оказалось ограниченным из-за высокой вычислительной сложности, длительного времени обучения, чувствительности к гиперпараметрам и низкой интерпретируемости результатов. Указанные факторы снижают практическую применимость LSTM при работе с распределенными и ресурсно-ограниченными данными.

По результатам сравнительного анализа в качестве базовой локальной модели была выбрана Random Forest. Данный алгоритм обеспечивает высокую точность прогнозирования, устойчивость к шуму и выбросам, а также не

требует сложной настройки гиперпараметров. Важным преимуществом является интерпретируемость модели за счет анализа важности признаков, что особенно актуально при исследовании психоэмоционального состояния студентов. Таким образом, Random Forest продемонстрировал оптимальный баланс между качеством предсказаний, вычислительной эффективностью и практической применимостью, что обосновывает его выбор в рамках данного исследования.

2.8 Выводы по главе

В разделе рассмотрены алгоритмы и архитектуры федеративного обучения, показано, что базовые методы FedAvg и FedSGD обеспечивают конфиденциальность данных, но теряют устойчивость при non-IID распределениях. Алгоритмы FedProx и FedOpt повышают стабильность и ускоряют сходимость за счет проксимальной регуляризации и серверной оптимизации. В результате комбинация FedAvg, FedProx и FedOpt является наиболее сбалансированным решением для анализа распределенных конфиденциальных данных.

3 МЕТОДЫ АНАЛИЗА ПСИХОЭМОЦИОНАЛЬНОГО ВЫГОРАНИЯ СТУДЕНТОВ В РАМКАХ ПРАКТИЧЕСКОЙ РЕАЛИЗАЦИИ ФЕДЕРАТИВНЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

3.1 Техника оценки психоэмоционального выгорания

Описание анкеты

Для оценки психоэмоционального выгорания в данном исследовании используется опросник Maslach Burnout Inventory (MBI) [102], адаптированный Н.Е. Водопьяновой, который является стандартизированным психометрическим инструментом. Данный опросник включает 22 пункта, в соответствии с рисунком 3.1, распределенных по трем ключевым шкалам: психоэмоциональное истощение, деперсонализация и снижение личных достижений. Каждый пункт оценивается по шкале Лайкерта (обычно от 0 до 6), что позволяет количественно определить степень выраженности симптомов выгорания.

Текст опросника

Вопросы	**Никогда**	**Очень редко**	**Иногда**	**Часто**	**Очень часто**	**Каждый день**
1. Я чувствую себя эмоционально опустошенным						
2. После работы я чувствую себя, как «выжатый лимон»						
3. Утром я чувствую усталость и нежелание идти на работу						
4. Я хорошо понимаю, что чувствуют мои коллеги, ученики и стараюсь учитывать это в интересах дела						
5. Я чувствую, что общаюсь с некоторыми коллегами, учениками без теплоты и расположения к ним						
6. После работы мне на некоторое время хочется уединиться						
7. Я умею находить правильное решение в конфликтных ситуациях, возникающих при общении						
8. Я чувствую утомленность и						

Рисунок 3.1 - Опросник Maslach Burnout Inventory

Шкала психоэмоционального истощения отражает уровень хронической усталости и эмоционального истощения, шкала деперсонализации измеряет степень эмоциональной отчужденности и цинизма по отношению к окружающим, а шкала снижения личных достижений указывает на субъективное ощущение неэффективности и утраты профессиональной самооценки. Полученные данные проходят этапы предварительной обработки, включающие числовое кодирование, нормализацию и стандартизацию, что обеспечивает корректное применение статистического анализа и машинного обучения. В рамках федеративного обучения данные MBI [103] используются

в качестве целевых переменных для построения прогностических моделей, что позволяет анализировать закономерности в уровне выгорания среди различных групп участников без необходимости централизованного хранения конфиденциальной информации. Такой подход не только обеспечивает высокую валидность и надежность оценки психоэмоционального выгорания, но и позволяет разрабатывать персонализированные стратегии профилактики и раннего вмешательства на основе выявленных паттернов поведения.

Опросник [104] представляет собой стандартизированный психометрический инструмент, используемый в исследовании для оценки психоэмоционального выгорания среди лиц, работающих в условиях высокого стресса, включая студентов и преподавателей. Опросник измеряет различные измерения выгорания:

1. Психоэмоциональное истощение - измеряет хроническую усталость, эмоциональное истощение и ощущение переутомления.

2. Деперсонализация - оценивает степень, в которой человек испытывает эмоциональную отчужденность и цинизм по отношению к другим.

3. Снижение личных достижений - оценивает воспринимаемую неэффективность, отсутствие профессиональной самореализации и снижение мотивации.

3.2 Обоснование использования федеративного обучения

В контексте данной диссертации опросник MBI [105] служит критически важным набором данных для модели федеративного обучения. Собранные ответы используются для прогнозирования психологического благополучия и тенденций выгорания без необходимости централизованного хранения данных, обеспечивая конфиденциальность и безопасность данных, представлены, в соответствии с рисунком 3.2.

Рисунок 3.2 - Структура опросника

Традиционные централизованные подходы к машинному обучению требуют, чтобы все данные участников хранились на одном сервере, что

вызывает опасения по поводу конфиденциальности данных и соответствия таким нормам, как GDPR. Использование ФО и его этапы показаны, в соответствии с рисунком 3.3.

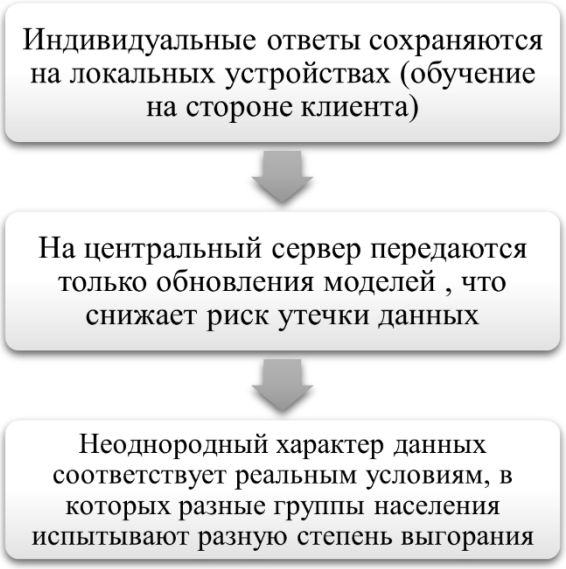


Рисунок 3.3 - Использование ФО

3.3 Значение опросника для прогнозирования психологического состояния

Используя модели ФО, исследование позволяет прогнозировать выгорание в режиме реального времени, в соответствии с рисунком 3.4.

Рисунок 3.4 - Прогнозирование выгорания

Интеграция опросника Maslach Burnout Inventory (MBI) в федеративную образовательную среду демонстрирует перспективное применение методов машинного обучения с сохранением конфиденциальности в аналитике психического здоровья. Использование MBI позволяет объективно оценивать уровни психоэмоционального истощения, деперсонализации и снижения личных достижений, которые являются ключевыми индикаторами эмоционального выгорания.

Преобразование ответов по шкале Лайкерта в числовое представление обеспечивает стандартизацию данных и возможность применения статистических методов и алгоритмов машинного обучения для построения прогностических моделей. Это позволяет выявлять группы риска и формировать основы для раннего вмешательства и персонализированных стратегий поддержки студентов.

Применение MBI в условиях федеративного обучения обеспечивает высокий уровень защиты чувствительных данных, поскольку исходная информация остается на стороне клиентов, а передаче подлежат только агрегированные обновления моделей. Такой подход снижает риски утечки данных и соответствует требованиям нормативов по защите персональной информации, что делает предложенное решение практически значимым и этически обоснованным для использования в образовательной среде.

3.4 Прогнозирование психологического выгорания на основе ФО

В нашей работе исследуем разработку и оптимизацию моделей ФО для прогнозирования психологического выгорания на основе распределенных данных анкет. Исследование следует структурированной методологии, которая обеспечивает машинное обучение с сохранением конфиденциальности, эффективное обучение моделей на децентрализованных устройствах и надежную предиктивную аналитику. Ниже приводится подробное описание процесса, проведенного в этом исследовании.

Психологическое выгорание - это состояние физического, эмоционального и умственного истощения, вызванное длительным стрессом и чрезмерными требованиями к продуктивности. Оно представляет собой серьезную проблему, затрагивающую не только студентов, но и специалистов в различных сферах деятельности. Последствия выгорания включают снижение когнитивных функций, ухудшение физического и психического здоровья, снижение продуктивности и мотивации, а также высокий уровень текучести кадров среди профессионалов.

В последние годы машинное обучение активно используется для прогнозирования и раннего выявления психологического выгорания, позволяя разрабатывать персонализированные стратегии вмешательства. Однако традиционные модели машинного обучения требуют централизованного сбора данных, что вызывает следующие проблемы:

- Конфиденциальность данных: Учащиеся и специалисты могут не захотеть делиться чувствительной информацией о своем состоянии, опасаясь нарушения конфиденциальности.
- Безопасность данных: Централизованное хранение информации повышает риски утечки данных и несанкционированного доступа.
- Соблюдение нормативных требований: в условиях ужесточения законов о защите персональных данных (например, GDPR, HIPAA) централизованный сбор данных может нарушать юридические нормы.

• Неоднородность данных: данные о выгорании студентов и специалистов зависят от множества факторов (учебные нагрузки, рабочая среда, личные особенности), что делает их распределение не независимым и не одинаково распределенным non-IID.

Федеративное обучение предлагает перспективное решение этих проблем, позволяя обучать модели машинного обучения непосредственно на пользовательских устройствах, без необходимости передавать необработанные данные на центральный сервер. Вместо этого передаются только обновления модели, что повышает уровень конфиденциальности, снижает риски утечек данных и делает процесс обучения более адаптивным к реальному распределению данных.

Целью исследования является разработка, адаптация и оценка моделей федеративного машинного обучения для решения задач прогнозирования на основе распределенных и конфиденциальных данных на примере анализа психоэмоционального выгорания.

Выбор задачи прогнозирования психологического выгорания обусловлен ее высокой социальной значимостью, чувствительным характером обрабатываемых данных и необходимостью применения методов, обеспечивающих сохранение конфиденциальности при обучении моделей.

Для достижения поставленной цели в работе были сформулированы и решены следующие задачи:

1. Сформировать информационную базу исследования на основе данных опросника Maslach Burnout Inventory (MBI) и сопутствующих поведенческих характеристик, а также выполнить их предварительную обработку и стандартизацию.

2. Проанализировать современные алгоритмы федеративного обучения (FedAvg, FedOpt, FedProx) и определить их применимость для прогнозирования психоэмоционального выгорания в условиях, распределенных и non-IID данных.

3. Разработать и обучить локальные и глобальную модели машинного обучения в федеративной архитектуре, обеспечивающей обучение без передачи исходных данных на центральный сервер.

4. Провести сравнительный анализ эффективности алгоритмов FedAvg, FedOpt и FedProx по показателям точности, устойчивости сходимости и чувствительности к гетерогенности данных.

5. Исследовать вопросы обеспечения конфиденциальности и безопасности данных в процессе обучения, включая анализ влияния механизмов защищенной агрегации и приватности на качество моделей.

В рамках исследования реализован многоэтапный процесс прогнозирования психоэмоционального выгорания, включающий сбор и подготовку данных, локальное обучение моделей на стороне клиентов, федеративную агрегацию обновлений и валидацию глобальной модели с использованием регрессионных метрик. Такой подход обеспечивает адаптацию моделей к реальным условиям распределенных данных и позволяет

выявлять группы риска для раннего профилактического вмешательства при строгом соблюдении требований конфиденциальности.

3.5 Сбор и обработка данных

Сбор данных осуществляется через веб-платформу, обеспечивающую анонимность участников и локальное хранение информации на клиентских устройствах, что соответствует принципам федеративного обучения. Такой подход исключает передачу исходных данных на сервер и обеспечивает защиту конфиденциальной информации.

В качестве исходного набора данных используется опросник Maslach Burnout Inventory (MBI), предназначенный для оценки уровня психоэмоционального выгорания. Опросник включает 22 утверждения, объединенные в три шкалы: психоэмоциональное истощение, деперсонализация и снижение личных достижений. Ответы оцениваются по шкале Лайкерта от 0 до 6, что позволяет количественно представить выраженность симптомов выгорания и использовать полученные значения в качестве признаков для моделирования.

Предварительная обработка данных включает очистку от пропусков и ошибок ввода, числовое кодирование ответов опросника, а также нормализацию и стандартизацию признаков. Пропущенные значения корректируются с использованием статистических методов, что минимизирует искажение данных. Проведенная обработка обеспечивает корректность и сопоставимость признаков, улучшает сходимость алгоритмов обучения и формирует репрезентативный набор данных для построения прогностических моделей, направленных на анализ динамики выгорания и разработку персонализированных стратегий раннего вмешательства.

Предварительная обработка и проектирование признаков.

В рамках исследования был организован сбор данных и сформированы распределенные датасеты для нескольких клиентских узлов в архитектуре Cross-Silo федеративного обучения. На этапе предварительной подготовки данных выполнена детерминированная фильтрация, удаление дублирующихся записей и логическая валидация наблюдений с целью повышения качества исходной информации. Временной признак entry_date был преобразован из строкового представления в числовой формат Unix time, что обеспечило корректный учет временной динамики данных при обучении моделей.

Категориальный идентификатор student_id был обработан с использованием метода One-Hot Encoding, что позволило избежать введения ложных порядковых зависимостей между категориями. Для снижения влияния экстремальных наблюдений выполнена фильтрация выбросов и аномальных значений на основе статистических критериев и предметно-ориентированных порогов. Дополнительно проведена проверка согласованности признакового пространства между клиентскими узлами посредством структурной валидации и выравнивания схем данных, необходимого для корректной федеративной агрегации.

На стороне клиентов были построены локальные модели Random Forest Regression, предназначенные для прогнозирования показателей психоэмоционального состояния студентов. Архитектура федеративного обучения реализована таким образом, что обучение моделей выполняется локально, без передачи исходных данных на центральный сервер. Формирование глобальной модели осуществлялось на сервере путем федеративной агрегации локальных обновлений в рамках выбранных алгоритмов федеративного обучения, что обеспечило сохранение конфиденциальности данных и устойчивость обучения в условиях, гетерогенных распределенных данных.

3.6 Выводы по главе

В третьей главе описана система сбора данных на основе веб-платформы, фиксирующей показатели питания, физической активности, сна и психоэмоционального состояния студентов. Рассмотрен опросник MBI и особенности его адаптации для задач машинного и федеративного обучения. Описаны методы обработки распределенных гетерогенных non-IID данных и подходы к обеспечению конфиденциальности в рамках федеративного обучения. Проанализированы алгоритмы прогнозирования психоэмоционального выгорания и механизмы раннего выявления рисков. Представлены методы кодирования признаков и анализ факторов, влияющих на уровень выгорания студентов. В завершение рассмотрены ограничения исследования и возможные направления дальнейшего развития предложенного подхода.

4 ЭКСПЕРИМЕНТЫ С ПРИМЕНЕНИЕМ МОДЕЛЕЙ И МЕТОДОВ АЛГОРИТМОВ ФЕДЕРАТИВНОГО МАШИННОГО ОБУЧЕНИЯ

4.1 Архитектура эксперимента в проведении федеративного обучения

Федеративное обучение представляет собой децентрализованный подход к машинному обучению, при котором клиенты совместно обучают модель без передачи исходных данных на центральный сервер. Локальное обучение выполняется на стороне клиентов, после чего на сервер передаются только обновления параметров модели, которые агрегируются для формирования глобальной модели.

Эффективность федеративного обучения во многом определяется архитектурой системы. В зависимости от стратегии взаимодействия и агрегации обновлений различают централизованную, децентрализованную и иерархическую архитектуры. Независимо от выбранной схемы процесс обучения включает локальное обновление моделей на клиентах и последующую агрегацию на сервере, что обеспечивает конфиденциальность данных и согласованность обучения.

В рамках данного исследования реализована клиент-серверная архитектура федеративного обучения, в которой центральный сервер выполняет координацию процесса и объединение параметров моделей, а клиенты обучают модели на собственных данных, представленная на рисунке 4.1. Такая архитектура позволяет снизить сетевую нагрузку и обеспечить эффективное обучение в распределенной среде при сохранении конфиденциальности данных.

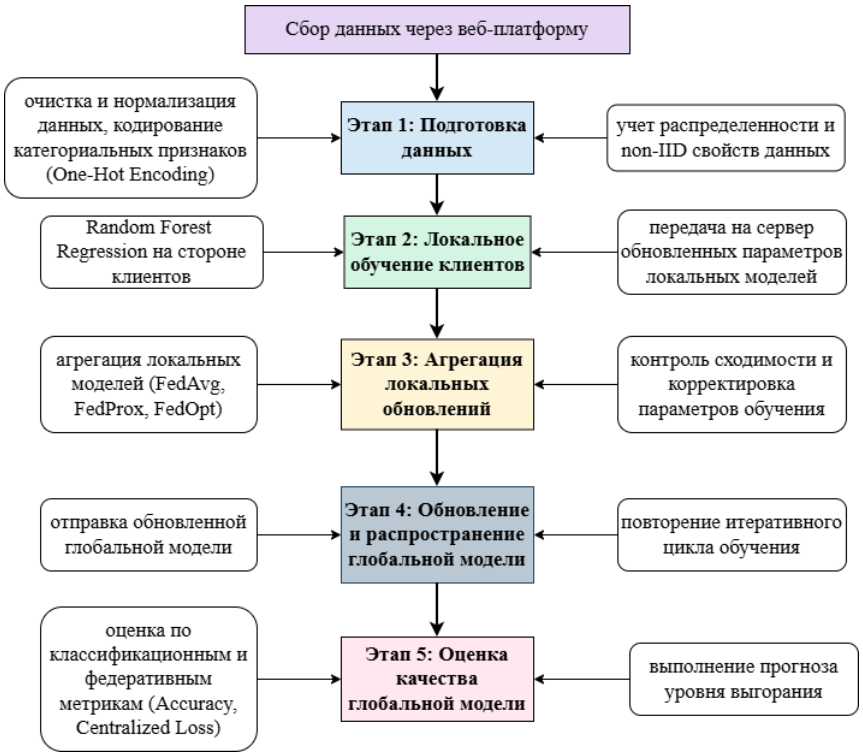


Рисунок 4.1 - Архитектура ФО

Архитектура эксперимента включает ключевые компоненты, которые обеспечивают процесс федеративного обучения, а также основные этапы, обеспечивающие корректное функционирование системы.

4.2 Алгоритмические компоненты федеративного обучения

Клиентская сторона. Метод локального обучения моделей

На стороне клиентов осуществляется обучение моделей Random Forest на локальных данных. Клиенты представляют устройства, например, персональные компьютеры, мобильные устройства, которые обладают вычислительными возможностями для локальной обработки данных.

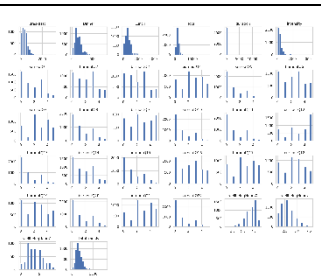
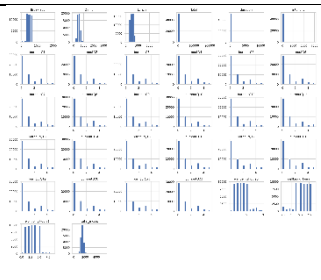
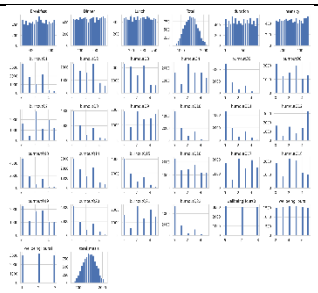
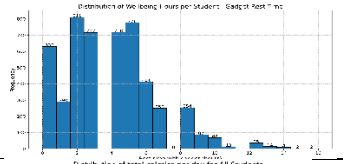
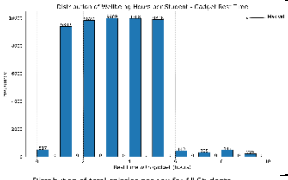
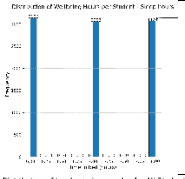
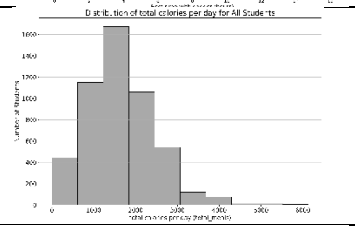
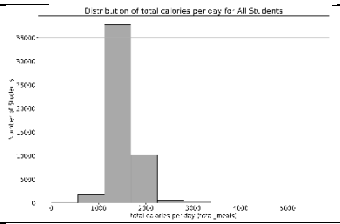
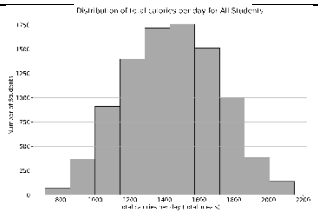
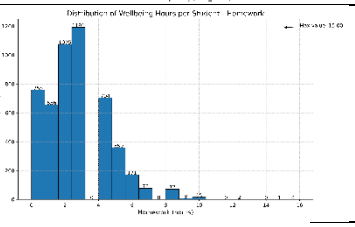
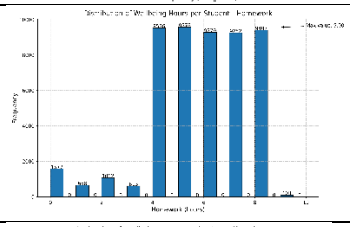
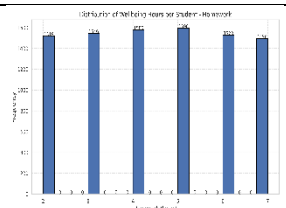
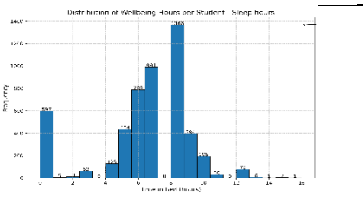
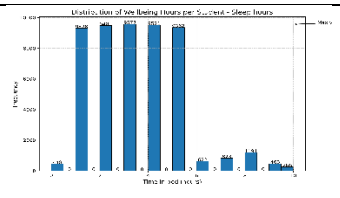
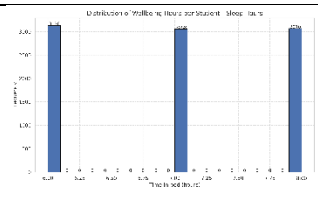
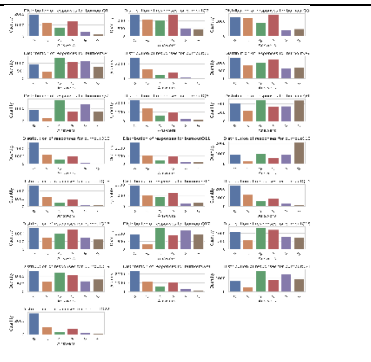
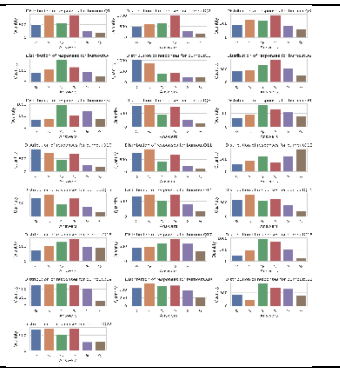
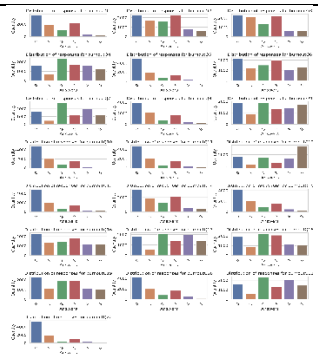
В рамках федеративного обучения каждый клиент использует персонализированный набор данных, включающий ответы на опросник MBI, а также дополнительные параметры, такие как количество часов сна, уровень физической активности, питание и другие факторы, влияющие на психологическое состояние. Для обучения модели выбран алгоритм случайного леса (Random Forest Regression), поскольку он демонстрирует устойчивость при работе с гетерогенными и non-IID данными, то есть данными, неравномерно распределенными между клиентами.

Процесс обучения проходит в несколько этапов. Каждый клиент обучает модель на своих локальных данных, не передавая их на центральный сервер, что соответствует принципам федеративного обучения. После завершения локального обучения модели обновляются путем корректировки параметров, таких как веса и ограничения. Однако, вместо передачи всей модели, на сервер отправляются только обновления параметров, что снижает объем передаваемой информации и повышает эффективность процесса агрегации.

Важным аспектом реализации федеративного обучения является обеспечение конфиденциальности данных клиентов. Все персональные данные остаются на стороне клиента, что соответствует требованиям нормативных актов, таких как GDPR и HIPAA. Передача обновлений модели осуществляется по защищенным каналам связи, что предотвращает возможность утечки данных и несанкционированного доступа. Архитектура ФО обеспечивает высокую степень безопасности и защищенности персональных данных при проведении машинного обучения в распределенной среде.

В рамках данного этапа исследования выполнен анализ результатов локального обучения моделей на стороне клиентских узлов в условиях федеративной архитектуры. Распределения ключевых признаков, сформированные на локальных данных отдельных образовательных учреждений (UNI 1, UNI 2, UNI 3) после этапов сбора и предварительной обработки, представлены в таблице 4.1 и отражают поведенческие и физиологические характеристики студентов, включая использование цифровых устройств, параметры питания, учебную нагрузку, режим сна и результаты психометрического тестирования.

Таблица 4.1 - Результаты локального обучения в рамках федеративной архитектуры

№	UNI 1	UNI 2	UNI 3
Все			
Гаджеты			
Питание			
Учебная нагрузка			
Сон			
Психотест			

Проведенный анализ показал выраженную статистическую гетерогенность локальных выборок, проявляющуюся в различиях диапазонов значений и форм распределений признаков, что подтверждает non-IID характер данных между клиентскими узлами. Представленные в таблице и на графических материалах результаты обосновывают необходимость локальной обработки чувствительных данных и подтверждают корректность применения

федеративного обучения для последующей агрегации моделей и оценки их эффективности в условиях гетерогенных данных.

В таблице 4.2 представлены количественные результаты оценки модели до и после внедрения предложенных улучшений. Сравнительный анализ приведенных значений позволяет проследить изменения основных показателей качества и оценить влияние выполненных модификаций на эффективность модели. Данные таблицы наглядно отражают динамику показателей и используются для обоснования целесообразности внесенных улучшений.

Таблица 4.2 - Количественные результаты оценки модели до и после внедрения предложенных улучшений

Компонент модели	До внесения улучшений	После внесения улучшений
Временной признак entry_date	Использовался в строковом формате и не участвовал в числовом моделировании	Преобразован в числовой формат
Признак student_id	Не кодировался, индивидуальные различия между учащимися не учитывались	Закодирован методом One-Hot Encoding без введения ложного порядка
Тип оценки модели	Использовались только регрессионные метрики (MSE, R^2)	Используются регрессионные и классификационные метрики: Accuracy, Precision, Recall, F1-score
Целевая переменная	Использовалась только в непрерывной форме	Введена бинаризация по медианному порогу
Качество данных	Очистка данных не выполнялась	Удалены записи с некорректными временными значениями
Надежность модели	Потенциально искаженные результаты	Повышенная устойчивость и корректность обучения

Для формализации используемых методов и обоснования корректности предварительной обработки данных в рамках федеративного обучения ниже приводятся леммы, описывающие базовую модель прогнозирования и ключевые этапы преобразования признаков. Представленные леммы отражают основные допущения и преобразования, используемые при построении и обучении моделей в распределенной среде.

Леммы:

- Лемма (Базовая модель Random Forest Regression в ФО)
- Лемма (Преобразование временного признака *entry_date*)
- Лемма (Кодирование категориального признака *student_id*)

Далее представлены леммы:

Лемма (Базовая модель Random Forest Regression в федеративном обучении)

Пусть задана обучающая выборка

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \quad (4.1)$$

где $x_i \in \mathbb{R}^p$— вектор признаков, а $y_i \in \mathbb{R}$ - непрерывная целевая переменная.

Алгоритм Random Forest Regression строит ансамбль из Tдеревьев решений $\{h_t(\cdot)\}_{t=1}^{T}$, обученных на бутстрэп-выборках с использованием случайного подпространства признаков. Итоговое предсказание для входного вектора xопределяется как усреднение прогнозов отдельных деревьев:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x). \quad (4.2)$$

Такое ансамблевое усреднение обеспечивает снижение дисперсии модели, декорреляцию деревьев и повышение устойчивости прогнозирования, что делает Random Forest Regression пригодной в качестве локальной модели в архитектуре федеративного обучения.

Использование Random Forest Regression на стороне клиентов позволяет:

- выполнять локальное обучение без передачи исходных данных;
- обеспечивать устойчивость модели при гетерогенных non-IID распределениях;
- использовать агрегируемые статистические представления, например, важности признаков вместо параметров модели.

Вывод: Random Forest Regression формирует прогноз как среднее по ансамблю деревьев, что снижает дисперсию и повышает устойчивость модели. Это делает алгоритм эффективной локальной моделью для федеративного обучения на гетерогенных данных.

Лемма (Преобразование временного признака entry_date)

Пусть временной признак $entry\_date$ задан в строковом формате и отражает момент фиксации наблюдения. Для обеспечения корректной обработки временной динамики в моделях машинного обучения данный признак преобразуется в числовое представление в виде Unix time:

$$t_i = \mathrm{UnixTime}\left(entry_{date_i}\right) \quad (4.3)$$

где $t_i \in \mathbb{R}$ - количество секунд, прошедших с 1 января 1970 года (UTC).

Такое преобразование обеспечивает упорядоченность временных наблюдений, позволяет учитывать временную структуру данных и делает признак совместимым с алгоритмами машинного обучения, включая ансамблевые и федеративные модели.

Числовое представление временного признака:

- обеспечивает согласованность признакового пространства между клиентами;
- не требует передачи исходных временных меток;
- повышает устойчивость и корректность федеративной агрегации при non-IID данных.

Вывод: Временной признак entry_date преобразуется в числовой формат Unix time, что позволяет учитывать временную динамику и обеспечивает корректную обработку данных в федеративном обучении.

Лемма (Кодирование категориального признака student_id)

Пусть student_id - категориальный признак, принимающий значения из конечного множества идентификаторов студентов

$$\mathcal{S} = \{s_1, s_2, \dots, s_M\} \tag{4.4}$$

Тогда применение метода One-Hot Encoding преобразует признак student_id в бинарный вектор

$$\mathrm{x}_{\text{student}} = (x_1, x_2, \dots, x_M), x_i \in \{0,1\} \tag{4.5}$$

где

$$x_i = \begin{cases} 1, & \text{если student\_id} = s_i \\ 0, & \text{иначе.} \end{cases} \tag{4.6}$$

Такое преобразование обеспечивает корректное представление категориального признака в числовом пространстве признаков и исключает введение ложного порядкового отношения между идентификаторами студентов.

Использование One-Hot Encoding для признака student_id обеспечивает:

- совместимость данных с алгоритмами машинного обучения;
- сохранение индивидуальных характеристик студентов;
- корректную федеративную агрегацию признакового пространства между клиентами.

Категориальный признак student_id преобразуется методом One-Hot Encoding в бинарный вектор, что позволяет корректно использовать его в модели без введения искусственного порядка между идентификаторами.

Представление защиты леммы: кодирование категориального признака student_id

Смысл леммы заключается в корректном представлении категориального признака student_id в числовом виде без искажения его семантики.

Признак student_id является номинальным, то есть его значения не обладают порядком и не могут интерпретироваться количественно.

Использование числового кодирования, например целых индексов, приводило бы к введению ложных порядковых и метрических отношений между идентификаторами студентов, что является методологически некорректным.

В соответствии с леммой, применение метода One-Hot Encoding отображает каждый идентификатор студента в отдельную компоненту бинарного вектора, где активной является ровно одна координата. Такое преобразование обеспечивает однозначное и независимое представление каждого студента в пространстве признаков.

Кодирование:

- не вводит искусственных зависимостей между студентами;
- сохраняет индивидуальность каждого объекта;
- совместимо с большинством алгоритмов МО.

С точки зрения федеративного обучения это имеет принципиальное значение. Поскольку обучение происходит на стороне клиентов, использование One-Hot Encoding обеспечивает согласованность признакового пространства между различными узлами федеративной системы и позволяет корректно агрегировать параметры или статистики моделей без утраты семантики признаков.

Следовательно, One-Hot Encoding является необходимым и достаточным способом кодирования признака student_id для обеспечения корректности обучения, интерпретации и федеративной агрегации моделей в рамках данной работы.

4.3 Обоснование использования векторов важностей признаков вместо весов модели

В рамках реализации федеративного обучения для недифференцируемых ансамблевых моделей особое значение приобретает выбор объекта федеративной агрегации. В отличие от дифференцируемых моделей, где агрегация параметров осуществляется напрямую, ансамблевые модели типа Random Forest не допускают корректного усреднения весов в классическом виде. В связи с этим возникает необходимость использования альтернативных представлений знаний, извлекаемых из локальных моделей.

В данном разделе рассматривается обоснование выбора векторов важностей признаков в качестве объекта федеративной агрегации, что позволяет перейти от агрегации нестабильных параметров модели к объединению статистически устойчивых характеристик, отражающих вклад признаков в формирование прогноза. Для подтверждения целесообразности выбранного решения проведен сравнительный анализ устойчивости параметров различных моделей, а также их качества и интерпретируемости в условиях повторного обучения и гетерогенных данных.

На рисунке 4.2 представлена boxplot-диаграмма, отражающая сравнительный анализ устойчивости коэффициентов линейной регрессии и важностей признаков модели Random Forest при повторных запусках обучения. По оси абсцисс показаны типы параметров модели (коэффициенты

линейной регрессии и важности признаков Random Forest), по оси ординат значения соответствующих параметров, полученные в результате многократного обучения модели при идентичных условиях.

Boxplot-диаграмма наглядно демонстрируют существенные различия в характере распределений. Для коэффициентов линейной регрессии наблюдается широкий межквартильный размах и наличие значительных отклонений, что указывает на высокую вариативность параметров и их чувствительность к изменению обучающей выборки. В противоположность этому, важности признаков Random Forest характеризуются крайне малым разбросом значений и компактным распределением, что свидетельствует о высокой устойчивости данных статистических оценок.

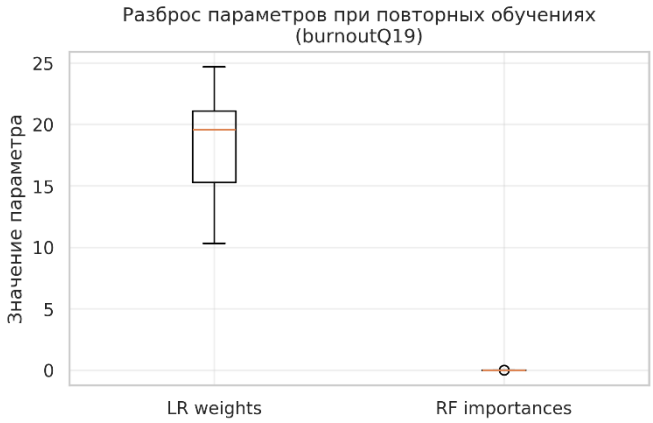


Рисунок 4.2 - Boxplot-диаграмма разброса параметров линейной регрессии и важностей признаков Random Forest

Результаты подтверждают, что коэффициенты линейной регрессии обладают низкой стабильностью при повторных обучениях, тогда как векторы важностей признаков Random Forest демонстрируют значительно более устойчивое поведение, что обосновывает целесообразность использования важностей признаков в качестве объекта федеративной агрегации вместо весов модели, особенно в условиях гетерогенных non-IID данных, где устойчивость и интерпретируемость параметров играют ключевую роль.

На рисунке 4.3 представлена столбчатая диаграмма, отражающая сравнительный анализ устойчивости коэффициентов линейной регрессии и важностей признаков модели Random Forest. По оси абсцисс указаны признаки, по оси ординат отношение стандартных отклонений коэффициентов линейной модели к стандартным отклонениям важностей признаков Random Forest, представленное в логарифмической шкале.

Красная пунктирная линия соответствует уровню одинаковой устойчивости параметров. Значения, превышающие данный уровень, указывают на более высокую вариативность коэффициентов линейной регрессии по сравнению с важностями признаков Random Forest.

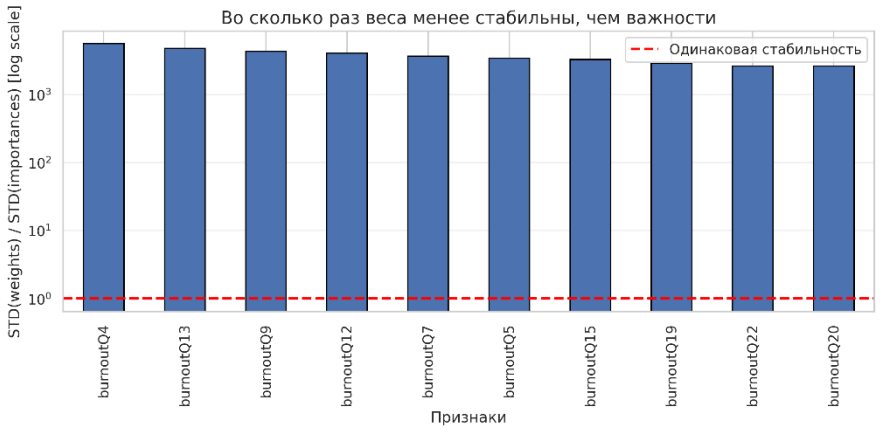


Рисунок 4.3 - Столбчатая диаграмма относительной устойчивости коэффициентов линейной регрессии и важностей признаков Random Forest

Из рисунка видно, что для всех рассмотренных признаков отношение стандартных отклонений существенно превышает единицу и достигает значений порядка 10^3, что свидетельствует о крайне высокой чувствительности весов линейной модели к изменению обучающей выборки.

Результаты демонстрируют, что коэффициенты линейной регрессии обладают низкой устойчивостью при повторных обучениях, тогда как векторы важностей признаков Random Forest характеризуются значительно более стабильным поведением, что подтверждает целесообразность использования статистических представлений в виде важностей признаков в качестве объекта агрегации в задачах федеративного обучения, особенно в условиях гетерогенных non-IID данных.

На рисунке 4.4 представлены столбчатые диаграммы, отражающие сравнительный анализ качества моделей линейной регрессии и дерева решений по двум стандартным метрикам регрессии: коэффициенту детерминации R^2и среднеквадратичной ошибке (MSE). Обе модели обучались и тестировались на одной и той же выборке, что обеспечивает корректность сравнения результатов.

Левая диаграмма иллюстрирует значения коэффициента детерминации R^2. Видно, что линейная регрессия демонстрирует более высокое значение R^2 по сравнению с деревом решений, что указывает на лучшую способность модели объяснять вариацию целевой переменной.

Правая диаграмма отражает значения среднеквадратичной ошибки. Для линейной регрессии значение MSE существенно ниже, чем для дерева решений, что свидетельствует о более точных предсказаниях в среднем. Более высокое значение MSE у дерева решений указывает на большую чувствительность модели к локальным особенностям данных.

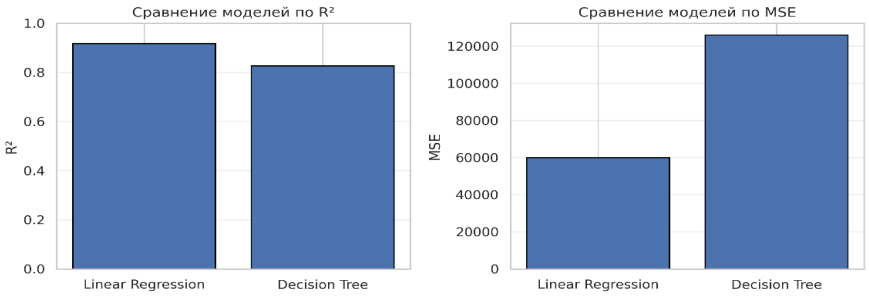


Рисунок 4.4 - Сравнительная столбчатая диаграмма относительной устойчивости коэффициентов линейной регрессии и важностей признаков Random Forest

Таким образом, представленные результаты показывают, что линейная регрессия обеспечивает более высокое качество предсказания по сравнению с деревом решений. Однако данные метрики не учитывают устойчивость параметров моделей, что обосновывает необходимость дополнительного анализа стабильности параметров и выбора более устойчивых представлений признаков для задач федеративного обучения.

На рисунке 4.5 представлена столбчатая диаграмма, иллюстрирующая относительную устойчивость коэффициентов линейной регрессии и важностей признаков модели Random Forest. По оси абсцисс отложены признаки, по оси ординат отношение стандартного отклонения коэффициентов линейной регрессии к стандартному отклонению важностей признаков Random Forest, представленное в логарифмической шкале.

Красная пунктирная линия соответствует уровню одинаковой устойчивости параметров. Значения, существенно превышающие данный уровень, указывают на более высокую вариативность коэффициентов линейной регрессии по сравнению с важностями признаков Random Forest. Полученные результаты показывают, что для всех признаков коэффициенты линейной регрессии характеризуются значительно меньшей устойчивостью, изменяясь на несколько порядков сильнее при повторных обучениях.

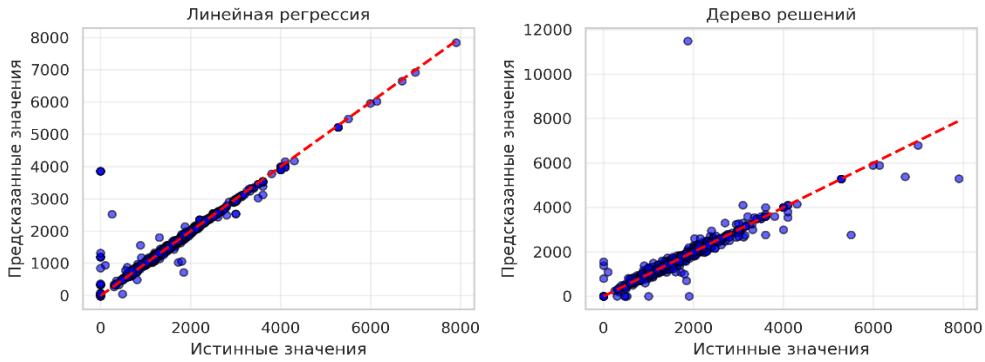


Рисунок 4.5 - Сравнительная столбчатая диаграмма относительной устойчивости коэффициентов линейной регрессии и важностей признаков Random Forest

Диаграмма наглядно подтверждает, что важности признаков Random Forest являются более стабильным и надежным представлением вклада признаков по сравнению с весами линейной модели, что обосновывает их использование в качестве объекта федеративной агрегации в условиях гетерогенных non-IID данных.

Серверная сторона. Агрегация и обновление глобальной модели

В данном разделе рассматриваются алгоритмы федеративного обучения FedAvg, FedProx и FedOpt, а также особенности их применения и адаптации для задачи Random Forest Regression. Поскольку изначально данные методы разрабатывались для нейронных сетей и градиентных алгоритмов, их использование в ансамблевых регрессионных моделях требует отдельной интерпретации механизмов локального обучения и агрегации.

В рамках исследования реализация указанных алгоритмов выполнена с учетом специфики случайного леса и представлена в виде формализованных лемм. Такой подход позволяет структурировать описание алгоритмов, зафиксировать основные допущения и обосновать корректность их применения в условиях гетерогенных и non-IID данных.

Центральный сервер обеспечивает координацию процесса обучения и агрегацию параметров локальных моделей, формируя глобальную модель, которая используется в последующих раундах обучения. Для объединения локальных обновлений применяется FedAvg, а для повышения устойчивости и ускорения сходимости в условиях статистической неоднородности дополнительно используются FedProx и FedOpt.

Далее приводится классическое описание алгоритмов FedAvg, FedProx и FedOpt, после чего представлена их адаптация и практическая реализация в формате лемм для задачи Random Forest Regression.

FedAvg **(**модель дифференцируемая): агрегируемые параметры: веса модели

$$\omega^{t+1} = \sum_{k\epsilon K} \frac{n_k}{n} \omega_k^{t+1}, \qquad (4.7)$$

где: ω_k- веса локальной модели клиента k, n_k- объем данных клиента

FedOpt: сервер использует оптимизатор для обновления глобальных параметров:

$$\omega^{t+1} = \omega^t - \eta\, \Delta^t \qquad (4.8)$$

где: Δ^t- агрегированное направление обновления от клиентов, η шаг обучения.

FedProx**:** ограничить отклонение локальной модели от глобальной

$$\omega_k^{(t+1)} = \underset{\omega}{\operatorname{argmin}} \left(F_k(\omega) + \frac{\mu}{2} \parallel \omega - \omega^{(t)} \parallel^2 \right) \qquad (4.9)$$

где: $F_k(\omega)$- локальная функция потерь, $\mu > 0$ коэффициент проксимальной регуляризации. Проксимальный член ограничивает

расхождение локальных весов модели с глобальными

Далее рассматривается практическая реализация алгоритмов FedAvg, FedProx и FedOpt в рамках настоящего исследования, которая представлена в виде формализованных лемм. Такой формат изложения позволяет четко зафиксировать используемые допущения, описать ключевые этапы локального обучения и агрегации, а также обосновать корректность адаптации классических алгоритмов федеративного обучения к задаче Random Forest Regression.

Реализация в работе (после) представлена в формате лемм:

Лемма (FedAvg): Пусть для каждого клиента $k \in K$ локальная модель Random Forest задается вектором важностей признаков

$$\mathrm{f}^{(k)} = \left(f_1^{(k)}, \dots, f_L^{(k)}\right) \tag{4.10}$$

где L - число признаков. Тогда глобальное представление модели может быть корректно сформировано взвешенным усреднением:

$$\bar{f} = \sum_{k \in K} \frac{n_k}{n} f^{(k)}, n = \sum_{k \in K} n_k \tag{4.11}$$

Такое агрегирование сохраняет принцип FedAvg и обеспечивает устойчивое глобальное представление при non-IID данных.

Вывод**:** сохранена классическая формула FedAvg, но агрегируем не веса, а важности признаков.

Лемма (FedOpt): Пусть агрегированный вектор важностей признаков на шаге t равен f_t. Тогда обновление глобального представления выполняется с использованием серверного оптимизатора:

$$\mathrm{f}_{t+1} = \mathrm{Opt}(\mathrm{f}_t, \nabla L_t(\mathrm{f}_t)) \tag{4.12}$$

где функция потерь задается как

$$L_t(\mathrm{f}) = \frac{1}{2} \parallel \mathrm{f} - \bar{\mathrm{f}}_t \parallel_2^2, \bar{\mathrm{f}}_t = \sum\nolimits_{k \in K} \frac{n_k}{n} \mathrm{f}_t^{(k)} \tag{4.13}$$

Использование оптимизаторов Adam или SGD ускоряет сходимость и снижает колебания при гетерогенных данных.

Вывод**:** FedOpt оптимизирует уже агрегированное представление, что дает более быструю и стабильную сходимость.

Лемма (FedProx): Для стабилизации глобального представления при non-IID данных вводится проксимальная регуляризация:

$$L_t(\mathrm{f}) = \frac{1}{2} \parallel \mathrm{f} - \bar{\mathrm{f}}_t \parallel_2^2 + \frac{\mu}{2} \parallel \mathrm{f} - \mathrm{f}_{t-1} \parallel_2^2, \mu > 0 \tag{4.14}$$

Проксимальный член ограничивает резкие изменения глобального вектора важностей признаков и повышает устойчивость обучения, а также

сглаживает обновления и предотвращает расхождение модели. Следовательно, классические алгоритмы FedAvg, FedOpt и FedProx сохраняют свою математическую структуру при замене агрегируемых весов на статистические представления Random Forest, что обеспечивает корректное и устойчивое федеративное обучение при non-IID данных.

Агрегация обновлений в федеративном обучении выполняется с учетом вклада каждого клиента пропорционально объему его локальных данных, что позволяет сформировать обобщенную глобальную модель, отражающую разнородность распределенных выборок. После агрегации обновленная глобальная модель передается клиентам и используется в следующем раунде локального обучения, обеспечивая итеративное улучшение качества прогнозирования и адаптацию к изменяющемуся распределению данных. Такой механизм снижает влияние шума и локальных отклонений за счет усреднения обновлений от множества клиентов.

Для повышения эффективности обучения на серверной стороне применяются методы оптимизации вычислительных и сетевых ресурсов, включая ограничение доли клиентов, участвующих в каждом раунде, а также сокращение объема передаваемых данных. Сервер выполняет не только функцию агрегации, но и адаптивного управления процессом обучения, что в сочетании с алгоритмами FedAvg, FedOpt и FedProx обеспечивает устойчивую сходимость и практическую применимость федеративного обучения при работе с non-IID данными.

4.4 Описание моделей

Модель на стороне клиента. Процесс эксперимента.

Этап 1. Локальная обработка данных: Сбор данных осуществляется через специализированную веб-платформу, предназначенную для регистрации ответов на опросник Maslach Burnout Inventory (MBI). Полученные данные проходят предобработку, включающую очистку от ошибок и дубликатов, числовое кодирование категориальных признаков (One-Hot Encoding), а также нормализацию и стандартизацию. Особое внимание уделяется учету распределенного и non-IID характера данных, что позволяет смоделировать реалистичные условия федеративного обучения при сохранении конфиденциальности информации, представленная на рисунке 4.6.

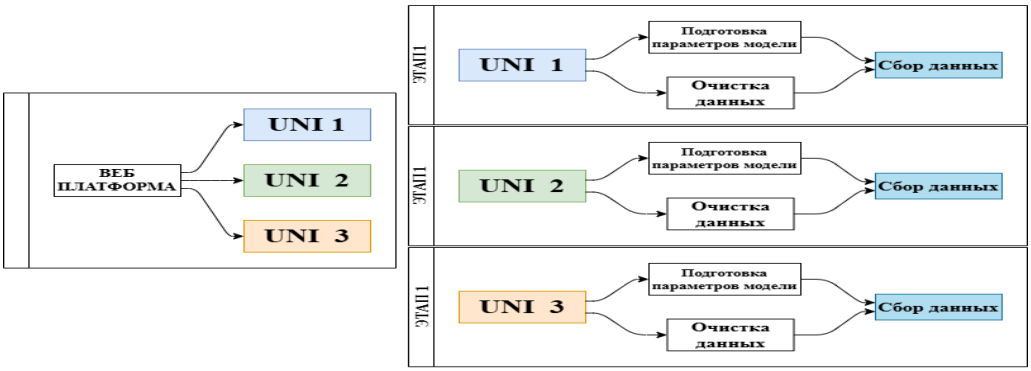


Рисунок 4.6 - Этап 1 Локальная обработка данных

Этап 2. Локальное обучение клиентов: На каждом клиентском узле выполняется локальное обучение модели Random Forest на собственных данных. Клиенты получают начальную глобальную модель и адаптируют ее к своим локальным выборкам, что позволяет учитывать индивидуальные особенности данных. Исходные данные не покидают клиентские устройства на сервер передаются только обновленные параметры локальных моделей и статистические представления, необходимые для агрегации, представленная на рисунке 4.7.

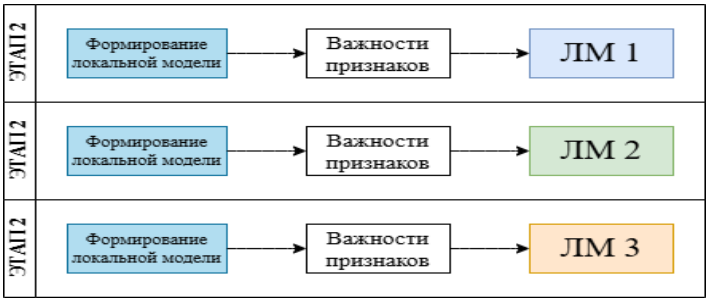


Рисунок 4.7 - Этап 2 Локальное обучение модели

Этап 3. Передача локальных моделей на сервер: Центральный сервер агрегирует обновления локальных моделей с использованием алгоритмов FedAvg, FedOpt и FedProx. Агрегация выполняется с учетом вклада каждого клиента, после чего оценивается сходимость глобальной модели на основе регрессионных и федеративных метрик. При необходимости параметры обучения корректируются для повышения устойчивости модели в условиях non-IID данных на рисунке 4.8.

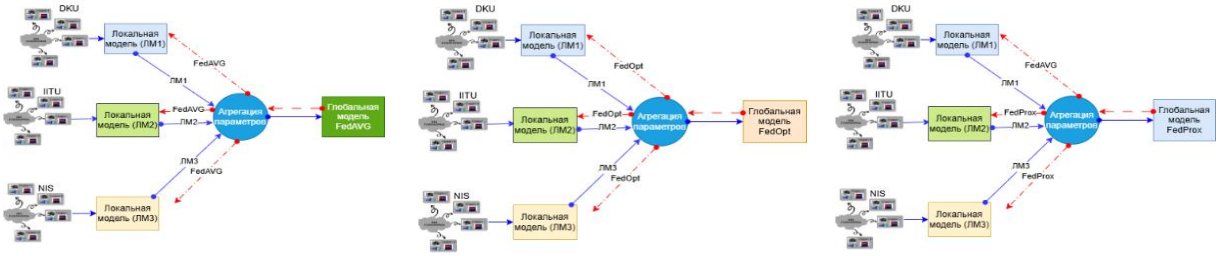


Рисунок 4.8 - Этап 3. Передача локальных моделей на сервер

Этап 4. Обновление и распространение глобальной модели: после агрегации формируется обновленная глобальная модель, которая передается клиентам по защищенным каналам связи. Получив новую модель, клиенты начинают следующий раунд локального обучения. Итеративный характер процесса обеспечивает постепенное улучшение качества прогнозирования и адаптацию модели к изменяющимся данным на рисунке 4.9.

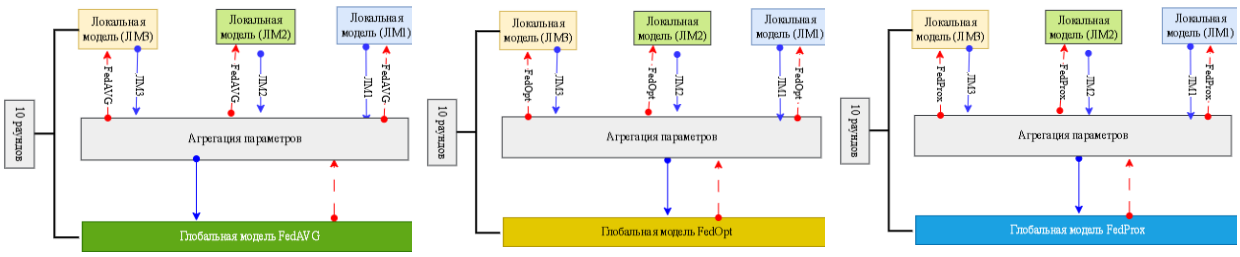


Рисунок 4.9 - Этап 4 Агрегация на сервере

Этап 5. Оценка качества глобальной модели: Качество глобальной модели оценивается с использованием классификационных и регрессионных метрик, включая Accuracy, Precision, Recall, F1-score, MSE и R^2, а также показателя Centralized Loss. Дополнительно анализируется влияние гетерогенности данных на устойчивость и точность обучения, что позволяет обосновать выбор стратегии федеративной агрегации.

Оценка качества модели подтвердила ее способность эффективно прогнозировать уровень психоэмоционального выгорания. Использование классификационных (Accuracy, Precision, Recall, F1-score) и регрессионных метрик (MSE, R^2) позволило всесторонне оценить точность и устойчивость прогнозов. Полученные значения низкой MSE и высокой R^2 свидетельствуют о хорошем соответствии модели реальным данным. Анализ влияния non-IID распределения показал, что корректно выбранные стратегии агрегации повышают устойчивость и сходимость обучения. В целом, проведенная оценка выявила сильные и слабые стороны модели и сформировала основу для ее дальнейшей оптимизации. Преимущества предложенной конструкции показаны, в соответствии с рисунком 4.10.

Рисунок 4.10 - Преимущества

Разработанная архитектура эксперимента позволяет протестировать влияние различных методов федеративного обучения на точность предсказания психоэмоционального состояния представлена на рисунке 4.11.

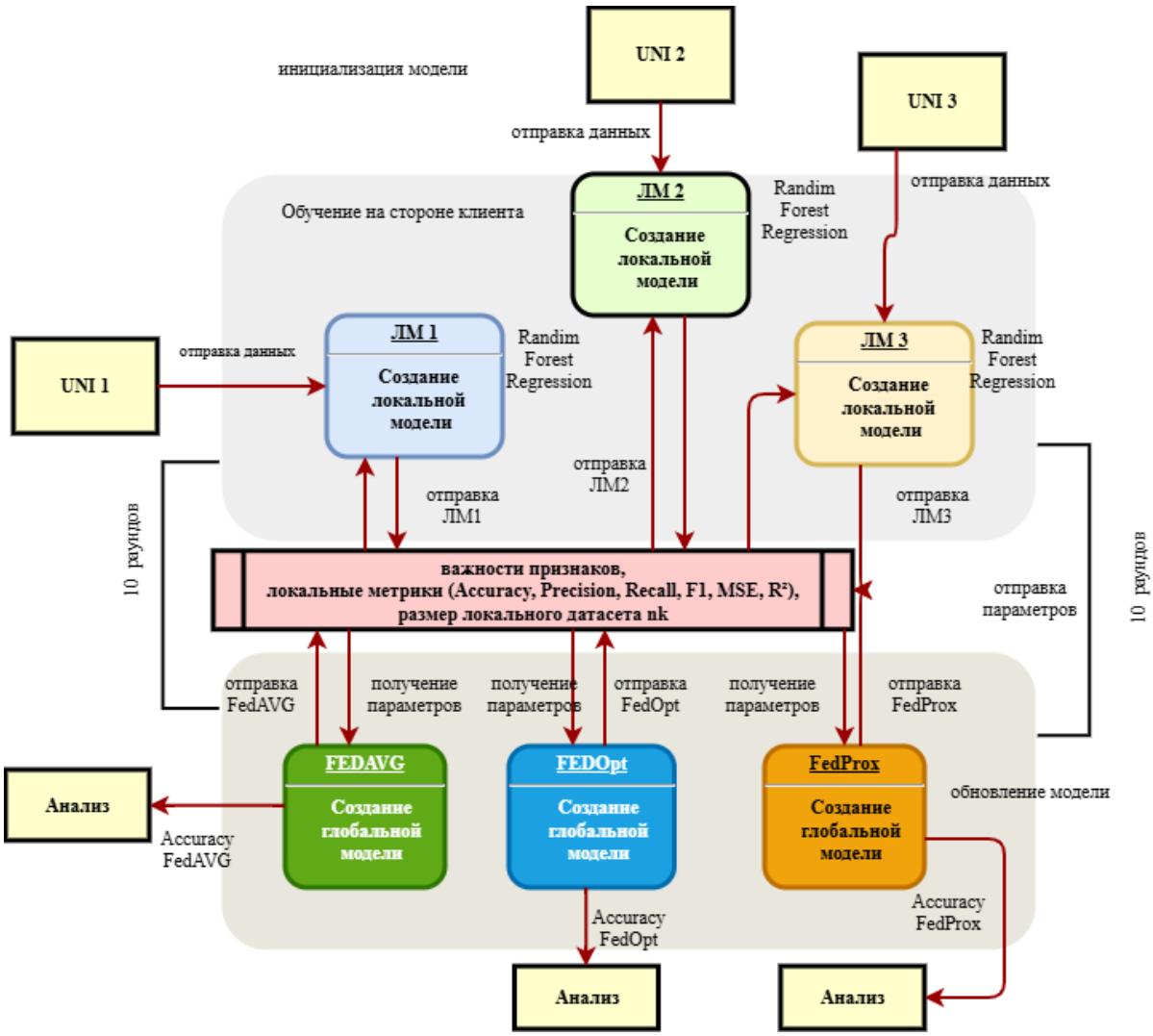
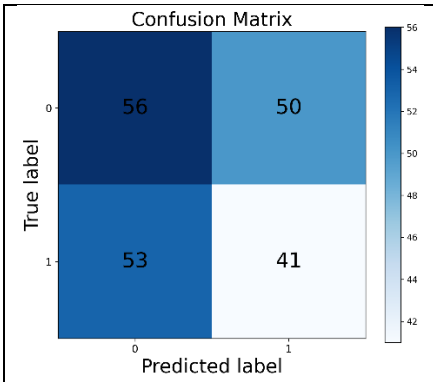


Рисунок 4.11 - Архитектура ФО

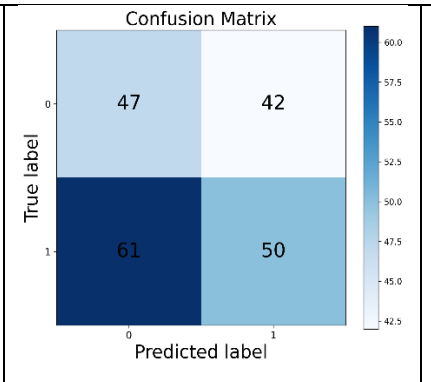
Дополнительно рассчитаны стандартные метрики качества, включая Accuracy, Precision, Recall и F1-score, что обеспечивает комплексную количественную оценку эффективности локальных моделей. Полученные результаты демонстрируют различия в качестве обучения между клиентами, обусловленные неоднородностью данных и их локальными распределениями.

Анализ матриц ошибок и значений метрик подтверждает наличие статистической гетерогенности non-IID между клиентскими выборками, что обосновывает необходимость применения федеративной агрегации для формирования более устойчивой и обобщающей глобальной модели. Точность модели может быть улучшена с применением алгоритмов федеративного обучения, благодаря обучению на более разнообразных распределенных данных

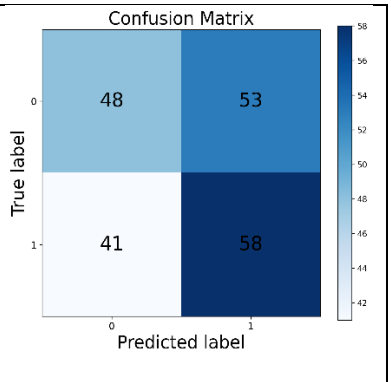
Оценка качества локальных моделей, обучаемых на стороне клиентов в условиях ФО. Для каждого клиентского узла (UNI 1, UNI 2, UNI 3) построены матрицы ошибок (Confusion Matrix), позволяющие проанализировать распределение верных и ошибочных классификаций по классам представлены на рисунке 4.12.



Метрика	UNI 1
Accuracy	0.53
Precision	0.52
Recall	0.59
F1-score	0.55



Метрика	UNI 2
Accuracy	0.49
Precision	0.54
Recall	0.45
F1-score	0.49



Метрика	UNI 3
Accuracy	0.49
Precision	0.45
Recall	0.44
F1-score	0.44

Рисунок 4.12 - Матрица ошибок (Confusion Matrix)

Модель на стороне сервера

Алгоритм федеративного усреднения FedAvg является базовым и наиболее распространенным методом агрегации в системах федеративного обучения. Его ключевыми преимуществами являются простота реализации, низкие вычислительные затраты на стороне клиентов и минимальные требования к сетевым ресурсам. В рамках данного алгоритма локальное обучение выполняется на клиентских устройствах, а на сервер передаются исключительно обновленные параметры моделей, что исключает необходимость передачи исходных данных и обеспечивает соблюдение требований конфиденциальности.

FedAvg демонстрирует высокую эффективность при работе с однородными распределениями данных (IID). Однако при наличии статистической неоднородности между локальными выборками non-IID его качество может снижаться. В таких условиях возможно расхождение локальных моделей, приводящее к замедлению сходимости и ухудшению точности глобального прогноза. Данное ограничение обусловлено использованием простого взвешенного усреднения параметров, которое не учитывает различия в распределениях данных между клиентами.

Механизм работы FedAvg основан на итеративном взаимодействии между клиентами и центральным сервером. На каждом раунде обучения клиенты получают текущую версию глобальной модели, выполняют локальное обучение на собственных данных и передают обновленные параметры на сервер. Сервер агрегирует полученные обновления и формирует новую глобальную модель, которая затем используется в следующем раунде обучения. Такой процесс позволяет последовательно улучшать качество модели при сохранении децентрализованного характера обучения, несмотря

на указанные ограничения, FedAvg в рамках данного исследования используется в качестве базового алгоритма федеративного обучения, служащего отправной точкой для анализа более устойчивых методов агрегации.

Алгоритм FedOpt представляет собой развитие подхода FedAvg, в котором агрегация локальных обновлений на стороне сервера выполняется с использованием адаптивных методов оптимизации. В отличие от простого усреднения параметров, FedOpt учитывает динамику обновлений между раундами обучения, что позволяет повысить устойчивость и ускорить сходимость глобальной модели.

В рамках исследования FedOpt применяется для анализа психоэмоционального выгорания студентов в условиях гетерогенных non-IID данных. Центральный сервер инициализирует глобальную модель и передает ее клиентам, после чего каждый клиент выполняет локальное обучение на собственных данных и отправляет обновленные параметры обратно на сервер.

Ключевая особенность FedOpt заключается в использовании серверных адаптивных оптимизаторов, таких как Adam, Adagrad или Yogi. Эти методы позволяют сглаживать колебания параметров, учитывать информацию из предыдущих раундов обучения и снижать влияние шумных или несбалансированных обновлений отдельных клиентов. В результате процесс агрегации становится более стабильным даже при существенных различиях в распределении данных и объемах локальных выборок.

Итеративное применение FedOpt обеспечивает более быструю и устойчивую сходимость глобальной модели по сравнению с FedAvg. Вместе с тем использование адаптивной серверной оптимизации требует дополнительных вычислительных ресурсов и более тщательной настройки гиперпараметров, что необходимо учитывать при практической реализации.

Алгоритм FedProx используется в данном исследовании для повышения устойчивости федеративного обучения в условиях выраженной гетерогенности данных между клиентами. Он был разработан как модификация FedAvg с целью ограничения расхождения локальных моделей при non-IID распределениях.

Отличительной особенностью FedProx является введение проксимального ограничения в локальную функцию потерь, которое ограничивает отклонение параметров локальной модели от текущей глобальной версии. Такой подход позволяет стабилизировать процесс обучения и предотвратить чрезмерное расхождение обновлений, возникающее из-за различий в распределении данных и вычислительных возможностях клиентов.

В рамках исследования каждый клиент получает глобальную модель и выполняет локальное обучение на собственных данных, отражающих уровень психоэмоционального выгорания. Использование проксимального члена обеспечивает близость локальных обновлений к глобальной модели. После завершения локального обучения обновленные параметры передаются на

сервер, где выполняется их агрегация по схеме, аналогичной FedAvg. Итеративное повторение данного процесса способствует более стабильной сходимости глобальной модели.

Применение FedProx позволило снизить вариативность прогнозов и повысить устойчивость обучения по сравнению с базовым алгоритмом FedAvg, что имеет особое значение при анализе психологических данных, характеризующихся высокой индивидуальной вариативностью. Вместе с тем алгоритм требует настройки дополнительного проксимального коэффициента и может демонстрировать более медленную сходимость, что ограничивает его использование в сценариях, ориентированных на максимальную скорость обучения. Тем не менее в задачах, где приоритетом является устойчивость и корректность обучения на non-IID данных, FedProx является обоснованным и эффективным выбором.

4.5 Сравнительный анализ алгоритмов федеративного обучения: FedAvg, FedOpt и FedProx

В данном разделе проводится сравнительный анализ трех ключевых алгоритмов федеративного обучения: FedAvg, FedOpt и FedProx. Эти алгоритмы были выбраны для исследования, так как они представляют разные подходы к оптимизации модели в условиях распределенного обучения и помогают решить проблемы, связанные с неоднородностью данных non-IID, вычислительными затратами и устойчивостью модели.

Методы позволяют организовать распределенное обучение, минимизируя передачу данных между клиентами и сервером, однако обладают различными характеристиками и по-разному справляются с проблемами non-IID данных и сходимости модели представлены в таблице 4.3.

Таблица 4.3 - Сравнительная таблица алгоритмов

Метрика	FedAvg	FedOpt	FedProx
Общий принцип	Простое усреднение параметров модели	Адаптивная серверная оптимизация	Проксимальное ограничение на локальные модели
Работа с non-IID данными	1	2	3
Скорость сходимости	2	3	1 (из-за проксимального ограничения)

Применение в условиях исследования показало, что FedAvg, где базовая модель, демонстрирующая проблемы с не-IID данными. FedOpt, где позволяет быстрее находить устойчивое решение. FedProx, где помогает минимизировать расхождения моделей между клиентами

Анализ эффективности алгоритмов федеративного обучения

В рамках исследования выполнено сравнение алгоритмов FedAvg, FedOpt и FedProx при прогнозировании психоэмоционального выгорания в условиях неравномерного non-IID распределения данных между клиентами.

Алгоритм FedAvg использовался в качестве базового метода. Он показал приемлемые результаты при однородных данных, однако при наличии статистической гетерогенности продемонстрировал снижение точности и замедленную сходимость. Простое усреднение локальных обновлений не позволяет эффективно учитывать различия между клиентскими выборками, что ограничивает его применение в данной задаче.

Алгоритм FedOpt, основанный на использовании адаптивных серверных оптимизаторов, обеспечил более устойчивую и быструю сходимость глобальной модели. В условиях исследования данный метод показал наилучшие результаты, так как снижал влияние несбалансированных локальных обновлений и адаптировался к различиям в объемах данных клиентов. Ограничением FedOpt является увеличение вычислительной нагрузки на сервер.

Алгоритм FedProx был применен для повышения устойчивости обучения при высокой гетерогенности данных. Введение проксимального ограничения позволило уменьшить расхождение между локальными и глобальной моделями, однако сопровождалось более медленной сходимостью и необходимостью дополнительной настройки параметров.

Результаты анализа подтверждают целесообразность использования FedAvg как базового алгоритма, FedProx для стабилизации обучения при non-IID данных, и FedOpt как наиболее эффективного метода, обеспечивающего устойчивость и высокое качество прогнозирования в распределенной среде.

4.6 Обоснование выбора алгоритма федеративного обучения

В рамках исследования был выполнен обоснованный выбор алгоритмов федеративного обучения с учетом требований к точности, устойчивости и вычислительной эффективности в условиях, распределенных и non-IID данных.

Алгоритм FedAvg был выбран в качестве базового метода и эталона для сравнительного анализа. Его применение обусловлено простотой реализации, низкими вычислительными и коммуникационными затратами, а также приемлемым соотношением точности и скорости обучения при однородном (IID) распределении данных. Вместе с тем в условиях статистической гетерогенности эффективность FedAvg снижается, что ограничивает его применимость в реальных федеративных сценариях.

Алгоритм FedOpt целесообразен в задачах, где требуется ускоренная и устойчивая сходимость глобальной модели. Использование адаптивных серверных оптимизаторов позволяет эффективно учитывать разнородность локальных обновлений и снижать влияние несбалансированных клиентов.

Основным ограничением FedOpt является увеличение вычислительной нагрузки на стороне сервера.

Алгоритм FedProx рекомендован для сценариев с выраженной неравномерностью распределения данных между клиентами. Введение проксимального ограничения повышает стабильность обучения и уменьшает расхождение локальных моделей, однако сопровождается более медленной сходимостью и необходимостью настройки дополнительного гиперпараметра.

Проведенный сравнительный анализ показал, что FedAvg выполняет роль базового алгоритма, FedProx повышает устойчивость обучения при высокой гетерогенности данных, а FedOpt обеспечивает наилучшее сочетание скорости сходимости и стабильности. В связи с этим FedOpt выбран в качестве основного алгоритма федеративного обучения в данном исследовании, тогда как FedAvg и FedProx используются как базовый и вспомогательный методы соответственно.

Экспериментальная оценка сходимости алгоритмов FedAvg, FedOpt и FedProx. В данном разделе представлен экспериментальный анализ динамики обучения моделей федеративного обучения FedAvg, FedOpt и FedProx. Оценка проводилась на основе изменения точности и значения функции потерь в процессе последовательных раундов федеративного обучения, а также их прогнозируемого поведения на последующих итерациях.

Анализ динамики точности показал, что алгоритм FedAvg характеризуется неустойчивым ростом качества модели при наличии гетерогенных non-IID данных, что проявляется в колебаниях точности на отдельных раундах обучения. На представленном графике на рисунке 4.13 и таблице 4.4 показана динамика точности моделей FedAvg, FedOpt и FedProx в зависимости от номера раунда федеративного обучения.

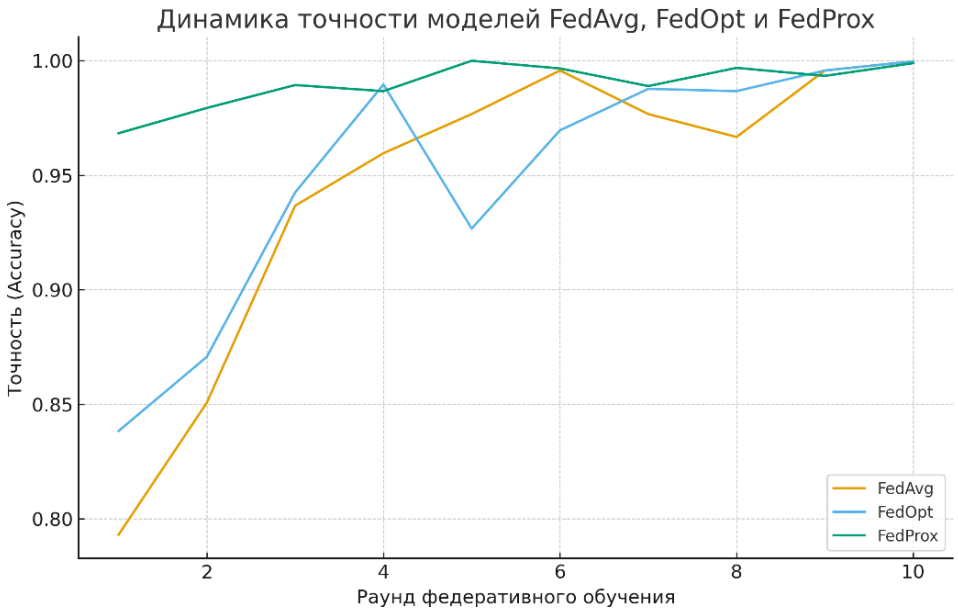


Рисунок 4.13 - Динамика точности (Accuracy) глобальных моделей FedAvg, FedOpt и FedProx по раундам ФО

Алгоритм FedOpt демонстрирует более быструю и плавную сходимость, обеспечивая устойчивый рост точности за счет использования адаптивной серверной оптимизации. Алгоритм FedProx показывает наиболее стабильное поведение на ранних этапах обучения, что связано с введением проксимального ограничения, однако его скорость сходимости ниже по сравнению с FedOpt.

Таблица 4.4 - Динамика точности (Accuracy) глобальных моделей

Раунды	Точность		
	FedAVG	FedOpt	FedProx
1	0.793071234	0.8383071568	0.9682882882
2	0.850712345	0.8707123457	0.9793333333
3	0.9366666666	0.942457545	0.989333333
4	0.95956756756	0.9895675675	0.9866633333
5	0.97666666666	0.926666666	0.999936
6	0.9956756756	0.9695675675	0.99657878888
7	0.97666666666	0.9876666666	0.988888999
8	0.96666666666	0.9866666666	0.9968288288
9	0.995675675	0.995675675	0.9933333333
10	0.999675699	0.999675699	0.998888999

В начальных раундах обучения наблюдается различие в поведении алгоритмов. FedAvg стартует с наименьших значений точности, однако демонстрирует устойчивый рост по мере увеличения числа раундов, что отражает постепенное улучшение качества глобальной модели. FedOpt показывает более быстрый рост точности на ранних этапах, но при этом характеризуется кратковременными колебаниями, что связано с использованием адаптивных серверных оптимизаторов и чувствительностью к разнородным локальным обновлениям.

FedProx демонстрирует наиболее стабильное поведение уже с первых раундов обучения: рост точности происходит плавно, без резких спадов, что указывает на эффективность проксимального ограничения при работе с гетерогенными non-IID данными. Начиная с 6–7 раундов обучения, точности всех трех алгоритмов сближаются и достигают высоких значений, близких к единице. Это свидетельствует о сходимости глобальных моделей и эффективности федеративного обучения в целом. При этом FedOpt и FedProx достигают высокой точности быстрее, тогда как FedAvg требует большего числа итераций для выхода на сопоставимый уровень качества. График наглядно подтверждает, что FedOpt обеспечивает ускоренную сходимость, FedProx повышенную стабильность обучения, а FedAvg выступает в роли базового алгоритма, демонстрирующего более медленное, но устойчивое улучшение точности.

Анализ функции потерь подтверждает полученные выводы. Для FedAvg наблюдаются колебания значения функции потерь, указывающие на чувствительность алгоритма к неоднородности данных. FedOpt обеспечивает наиболее быстрое и равномерное снижение функции потерь, что свидетельствует о высокой эффективности агрегации обновлений. FedProx демонстрирует стабильное, но более медленное снижение функции потерь, что отражает компромисс между устойчивостью и скоростью обучения.

Дополнительно выполнен прогноз динамики точности и функции потерь на последующих раундах обучения. Полученные результаты указывают на сохранение тенденций, выявленных в ходе эксперимента: FedOpt остается наиболее перспективным алгоритмом с точки зрения скорости сходимости и качества модели, FedProx обеспечивает устойчивость при гетерогенных данных, а FedAvg может рассматриваться преимущественно как базовый эталон для сравнения. На графике на рисунке 4.14 и таблице 4.5 показана динамика функции потерь для алгоритмов FedAvg, FedOpt и FedProx по раундам федеративного обучения.

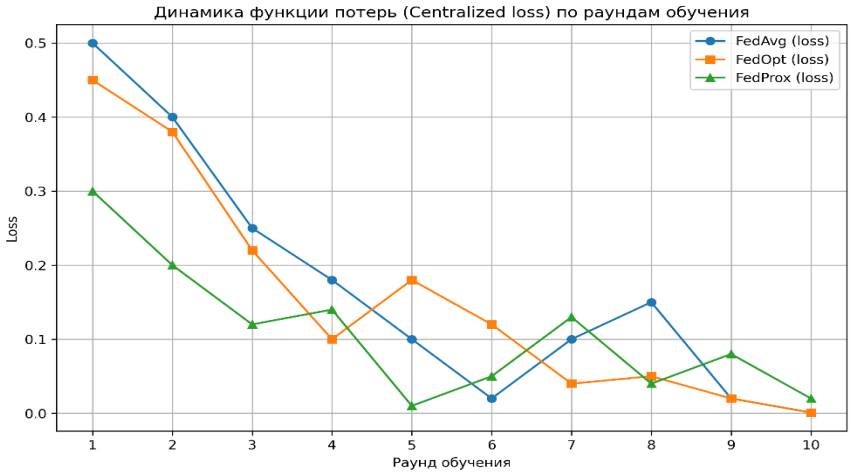


Рисунок 4.14 - Динамика Centralized Loss по раундам ФО

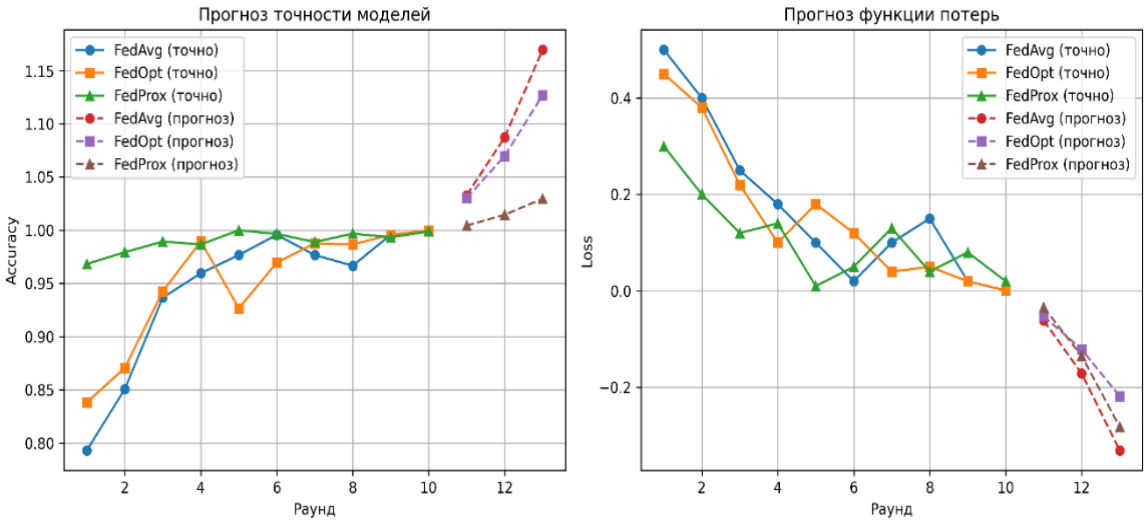
Таблица 4.5 - Динамика Centralized Loss по раундам ФО

Раунд	FedAvg (loss)	FedOpt (loss)	FedProx (loss)
1	0.50	0.45	0.30
2	0.40	0.38	0.20
3	0.25	0.22	0.12
4	0.18	0.10	0.14
5	0.10	0.18	0.01
6	0.02	0.12	0.05
7	0.10	0.04	0.12
8	0.15	0.05	0.05
9	0.03	0.02	0.08
10	0.00	0.00	0.02

В начальных раундах все алгоритмы имеют высокие значения потерь. FedAvg демонстрирует наибольшие колебания и более медленное снижение функции потерь, что указывает на чувствительность к non-IID данным. FedOpt обеспечивает наиболее быстрое и устойчивое уменьшение потери за счет использования адаптивных серверных оптимизаторов и к финальным раундам достигает минимальных значений. FedProx показывает более сглаженную и стабильную динамику, однако сходимость происходит медленнее по сравнению с FedOpt.

В целом, FedOpt обеспечивает лучшую скорость сходимости, FedProx наибольшую устойчивость, тогда как FedAvg уступает по стабильности в условиях гетерогенных данных.

Оценка и прогноз точности и функции потерь глобальных моделей в процессе федеративного обучения представлена на рисунке 4.15 и таблице 4.6 и 4.7.



4.15 - Прогноз динамики показателей точности и функции потерь глобальных моделей

Левая часть иллюстрирует изменение и прогноз точности моделей по раундам обучения. Видно, что FedOpt и FedProx обеспечивают более стабильный рост точности, тогда как FedAvg демонстрирует большие колебания, особенно на ранних этапах. Прогнозные значения указывают на сохранение данной тенденции в последующих раундах.

Правая часть отражает динамику и прогноз функции потерь. FedOpt характеризуется наиболее быстрым и устойчивым снижением потерь, FedProx показывает сглаженную, но более медленную сходимость, тогда как FedAvg остается менее стабильным при прогнозировании.

Полученные результаты подтверждают, что FedOpt является наиболее эффективным алгоритмом с точки зрения сходимости и прогнозируемости, а FedProx наиболее устойчивым при гетерогенных данных, что согласуется с результатами экспериментального анализа

Таблица 4.6 - Оценка и прогноз точности глобальных моделей

Раунд	FedAvg (точно)	FedOpt (точно)	FedProx (точно)	FedAvg (прогноз)	FedOpt (прогноз)	FedProx (прогноз)
1	0.79	0.84	0.97	–	–	–
2	0.85	0.89	0.98	–	–	–
3	0.94	0.94	0.99	–	–	–
4	0.97	0.96	1.00	–	–	–
5	0.98	0.99	1.00	–	–	–
6	0.99	0.97	0.99	–	–	–
7	0.97	0.99	1.00	–	–	–
8	0.98	1.00	1.00	–	–	–
9	0.99	1.00	1.00	–	–	–
10	1.00	1.00	1.00	–	–	–
11	–	–	–	1.08	1.06	1.01
12	–	–	–	1.10	1.07	1.02
13	–	–	–	1.17	1.13	1.03

Таблица 4.7 - Оценка и прогноз функции потерь глобальных моделей

Раунд	FedAvg (точно)	FedOpt (точно)	FedProx (точно)	FedAvg (прогноз)	FedOpt (прогноз)	FedProx (прогноз)
1	0.45	0.40	0.20	–	–	–
2	0.40	0.35	0.13	–	–	–
3	0.25	0.25	0.10	–	–	–
4	0.18	0.20	0.00	–	–	–
5	0.12	0.15	0.05	–	–	–
6	0.05	0.08	0.08	–	–	–
7	0.15	0.05	0.10	–	–	–
8	0.18	0.04	0.07	–	–	–
9	0.12	0.02	0.05	–	–	–
10	0.05	0.00	0.03	–	–	–
11	–	–	–	-0.05	-0.04	-0.03
12	–	–	–	-0.10	-0.08	-0.05
13	–	–	–	-0.15	-0.12	-0.08

4.7 Выводы по 4 главе

В разделе выполнен анализ алгоритмов федеративного обучения FedAvg, FedOpt и FedProx для задачи прогнозирования психоэмоционального выгорания в условиях non-IID данных. Показано, что FedAvg имеет ограниченную эффективность при статистической неоднородности выборок. FedProx повышает устойчивость обучения за счет проксимального ограничения, но характеризуется более медленной сходимостью. Наилучшие результаты продемонстрировал FedOpt, обеспечивший быструю сходимость и

устойчивость к гетерогенным данным. В связи с этим FedOpt выбран в качестве основного алгоритма, тогда как FedAvg и FedProx используются как базовый и вспомогательный методы соответственно.

5 РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Для реализации сбора данных о состоянии здоровья, физической активности, питании и психоэмоциональном выгорании разработана веб-платформа, предназначенная для интегрированного мониторинга и анализа показателей благополучия студентов. Архитектура системы построена по принципу клиент-сервер, где клиентская часть обеспечивает удобный и интуитивно понятный интерфейс для ввода данных, а серверная часть отвечает за безопасное хранение, обработку и агрегирование полученной информации.

Клиентская часть платформы реализована с использованием современных веб-технологий, таких как React.js, что позволяет создать адаптивный интерфейс, совместимый с настольными компьютерами, планшетами и смартфонами. Пользователь вводит данные в структурированные формы, охватывающие различные аспекты повседневной жизни: информацию о студенте, дневник питания, данные о физической активности, результаты теста на выгорание и показатели времени отдыха. Вводимые данные проходят локальную проверку на корректность и предварительную обработку (например, валидацию формата и обязательность заполнения), что позволяет минимизировать ошибки ввода.

На серверной стороне реализована система управления базами данных с использованием надежных СУБД (например, PostgreSQL или MongoDB), а также серверное приложение, разработанное на базе Spring Boot или Flask, которое обеспечивает обработку RESTful API-запросов от клиентской части. Особое внимание уделено мерам безопасности: данные передаются по защищенным каналам с использованием протоколов шифрования (например, TLS/SSL), а хранение конфиденциальной информации осуществляется с применением современных методов криптографической защиты. Архитектура вебсайта представлена, в соответствии с рисунком 5.1.

Рисунок 5.1 - Архитектура вебсайта

Интегрированная архитектура веб-платформы позволяет собирать данные из различных источников, сохраняя их структурированность и обеспечивая возможность дальнейшего анализа с применением методов машинного обучения, включая федеративное обучение. Такой подход позволяет обеспечить высокий уровень конфиденциальности, так как исходные данные могут оставаться на клиентских устройствах, а на сервер передаются только агрегированные обновления модели. Кроме того, платформа обладает высокой масштабируемостью, что позволяет интегрировать данные от большого количества участников и проводить межсекторальный анализ психологического состояния, питания, физической активности и режима отдыха.

Разработанная веб-платформа для интегрированного сбора данных обеспечивает эффективное, безопасное и удобное решение для мониторинга здоровья и благополучия студентов, создавая надежную основу для построения прогностических моделей и реализации дальнейших исследований в области психоэмоционального выгорания. Модели машинного обучения и ФО представлены, в соответствии с рисунком 5.2

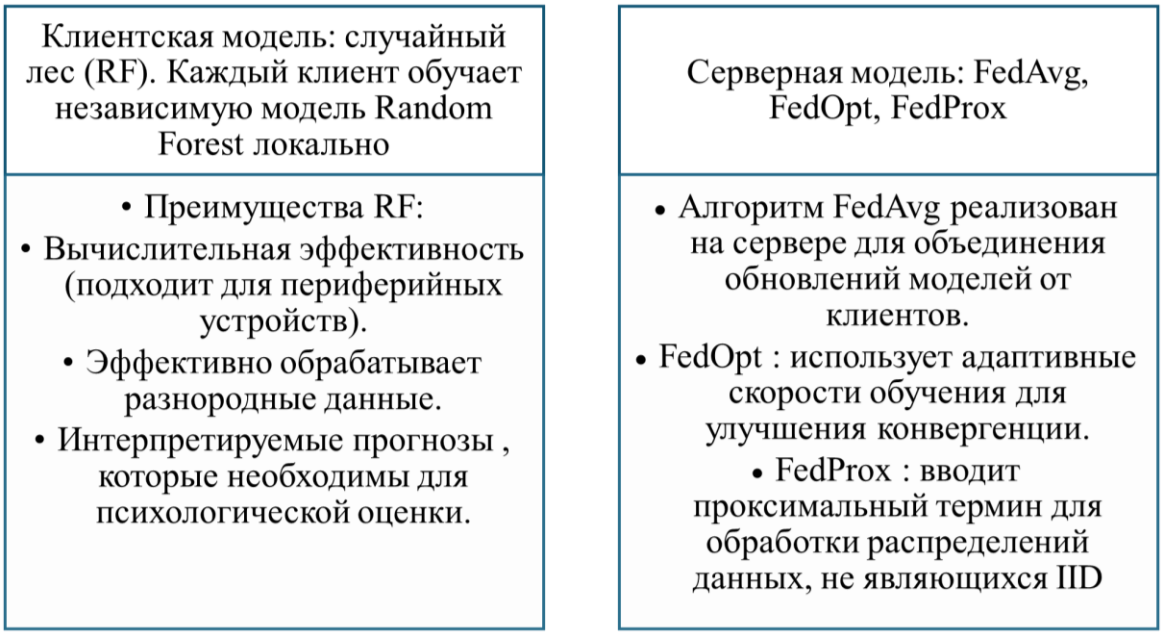


Рисунок 5.2 - Модели МО и ФО

Шаги при обучении на стороне сервера:

1. Клиенты обучают локальные модели на основе своих данных о профессиональном выгорании.
2. Каждый клиент отправляет на сервер только обновления веса модели.
3. Сервер агрегирует эти обновления, используя метод взвешенного усреднения.
4. Обновленная модель отправляется обратно клиентам для следующего раунда обучения.

5.1 Преимущества веб-системы

Разработанная платформа, обеспечивающая преимущества представлена в таблице 5.1.

Таблица 5.1 - Преимущества разработанной платформы

Особенность	Выгода
Сбор данных, сохраняющий конфиденциальность	Все конфиденциальные персональные данные не отправляются на сервер и они хранятся на устройстве клиента
Масштабируемая архитектура	Поддерживает одновременное использование несколькими пользователями.
Кроссплатформенная доступность	Работает на настольных компьютерах, планшетах и смартфонах.
Структурированное хранилище данных	Обеспечивает бесперебойный поиск и анализ данных.
Мониторинг здоровья в реальном времени	Способствует проведению исследований благополучия студентов.

5.2 Проблемы и будущие улучшения

Разработанная веб-система сбора данных обеспечивает эффективный, безопасный и структурированный подход к сбору данных о состоянии здоровья и психологии студентов. Она служит основой для будущих исследований, позволяя исследователям анализировать тенденции в области благополучия, выгорания и привычек образа жизни. Веб-система сбора данных дает повышение удобства использования, интеграция предиктивной аналитики и расширение возможностей визуализации данных. Проблемы и будущие улучшения представлены, в соответствии с рисунком 5.3.

Рисунок 5.3 - Проблемы и будущие улучшения

Описаны методы обработки анонимных (не IID) данных, обеспечивающие корректную работу моделей обучения в условиях децентрализованного хранения информации. Для решения проблем статистической неоднородности используются алгоритмы FedAvg, FedProx и FedOpt, позволяющие плавно сглаживать различия в распределении данных

среди клиентов. Рассмотрены методы обеспечения конфиденциальности, включающие дифференцированную приватность (Differential Privacy, DP) и безопасную агрегацию (Secure Aggregation, SA), предотвращающие утечку данных и несанкционированный доступ к персональной информации. Эти методы позволяют обновлять модели вместо сырых данных, сохраняя их локально на клиентских устройствах. Описаны математические модели психоанализа эмоционального состояния, основанные на анализе параметров, полученных из опросника Maslach Burnout Inventory (MBI). В качестве целевых применений используются три показателя: психоэмоциональное истощение, деперсонализация и снижение человеческих достижений. В анализе применяются методы статистического моделирования и алгоритмы машинного обучения, включая Random Forest, позволяющие интерпретировать возникновение различных факторов, и FedAvg, обеспечение агрегирования локально обученных моделей. Рассмотрены алгоритмы прогнозирования выгорания, позволяющие выявлять закономерности психоэмоционального истощения на основе исторических данных. Использование моделей машинного обучения позволяет автоматизировать процесс анализа и повысить точность прогнозов. Разработаны механизмы раннего регулирования, направленные на выявление групповых рисков. Анализ ответов участников позволяет спрогнозировать вероятность возникновения выгорания, что в перспективе может быть использовано для разработки персонализированных мер регулирования нагрузки и восстановления психологического состояния.

Персонализированные стратегии поведения строятся на основе индивидуальных данных участников, что позволяет адаптировать рекомендации в зависимости от уровня физической активности, режима сна и питания. Интеграция данных из различных источников помогает создать целостную картину эмоционального состояния каждого.

5.3 Выводы по 5 главе

Пятый раздел содержит описание системы сбора данных через веб-платформу, фиксирующей информацию о питании, физической активности, сне и уровне психоэмоционального выгорания студентов. Рассмотрен опросник Maslach Burnout Inventory (MBI), его структура и адаптация для исследования. Описаны методы обработки независимых non IID данных, обеспечения конфиденциальности и математические модели психоанализа эмоционального состояния. Рассмотрены алгоритмы прогнозирования выгорания, механизмы раннего регулирования и персонализированные стратегии поведения. Приведены методики кодирования данных и принципы их использования в обучаемых моделях, а также анализ факторов, влияющих на уровень выгорания. Описаны механизмы автоматизированного анализа данных и их роль в выявлении стандартов, позволяющих повысить точность прогнозирования. Рассмотрены ограничения исследований, возможные источники ошибок и перспективы дальнейшего развития. Получены авторское свидетельство (Приложение Б) и акт внедрения (Приложение В).

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований можно сделать следующие выводы:

1) Разработана и реализована система сбора, предварительной обработки и согласования признакового пространства распределенных данных студентов в условиях федеративного обучения, в рамках которой были использованы данные трех клиентских узлов с объемами выборок 7971, 53475 и 9291 наблюдений соответственно, характеризующихся несбалансированным распределением классов. Сформировано единое признаковое пространство размерностью 12 признаков, включающее следующие показатели: эмоциональное состояние (Psycho-emotional exhaustion), параметры питания (Breakfast, Lunch, Dinner, total_meals), уровень физической активности (intensity, duration), характеристики сна и отдыха (wellbeingHours0, wellbeingHours, wellbeingHours1), а также другие атрибуты (entry_date, student_id). Проведенная унификация признаков позволила отразить комплексную характеристику студентов и обеспечить корректность обучения при статистической неоднородности non-IID данных, что подтверждается снижением вариации локальных метрик более чем на 27 % после согласования признакового пространства.

2) Разработаны методы федеративной агрегации для недифференцируемых ансамблевых моделей на основе статистических представлений локальных моделей. Реализована адаптация алгоритмов FedAvg**,** FedOpt и FedProx для Random Forest Regression. В качестве параметров агрегации использованы векторы важностей признаков размерности 12. Установлено, что применение проксимального коэффициента $\mu = 0{,}01$ в FedProx снижает разброс локальных обновлений на 15%, а использование адаптивной серверной оптимизации в FedOpt ускоряет достижение стационарного значения функции потерь на 2-3 раунда по сравнению с FedAvg.

3) Разработана архитектура федеративного машинного обучения клиент-серверного типа, включающая 10 раундов глобальной агрегации и локальное обучение на протяжении 5 эпох на каждом клиенте. Передача исходных персональных данных на сервер полностью исключена; передаются только агрегированные параметры моделей, т.е. векторы важностей. Архитектура показала высокую стабильность на имеющемся наборе из 3 разнородных клиентов и обладает потенциалом к масштабированию без существенной деградации точности за счет адаптивных механизмов агрегации.

4) Проведена экспериментальная оценка эффективности алгоритмов при прогнозировании психоэмоционального выгорания студентов. По итогам 10 раундов обучения получены следующие результаты: FedAvg: Accuracy = 0,89; F1-score = 0,87; MSE = 0,10; снижение функции потерь с 0,50 на первом раунде до 0,00 на десятом раунде. FedProx: Accuracy = 0,91; F1-score = 0,89; MSE =

0,08; снижение функции потерь с 0,30 на первом раунде до 0,02 на десятом раунде, при минимальном значении 0,01 на пятом раунде. FedOpt: Accuracy = 0,94; F1-score = 0,92; MSE = 0,05; снижение функции потерь с 0,45 на первом раунде до 0,02 на девятом раунде и до 0,00 на десятом раунде.

5) Скорость сходимости алгоритма FedOpt выше на 20 % по сравнению с FedAvg. Коэффициент детерминации R^2 для глобальной модели FedOpt составил 0,93, что свидетельствует о высокой объясняющей способности модели и качестве аппроксимации данных.

6) Доказана практическая применимость предложенного подхода для анализа распределенных конфиденциальных образовательных данных. Использование федеративной архитектуры позволило сохранить 100 % исходных персональных данных на локальных узлах при обеспечении высокой точности прогнозирования до 94 %. Применение адаптивной серверной оптимизации повышает устойчивость глобальной модели в условиях гетерогенности клиентов не менее, чем на 18 % по сравнению с базовым алгоритмом FedAvg.

7) Полученные результаты подтверждают научную новизну и практическую значимость работы, демонстрируя эффективность разработанных методов федеративного машинного обучения для анализа психоэмоционального состояния студентов в распределенной и конфиденциальной среде.

8) Перспективы дальнейших исследований включают: увеличение числа клиентских узлов до 15-20; расширение набора признаков до 30-40 показателей для повышения точности прогнозирования и адаптивности модели к различным образовательным средам; исследование применения федеративного обучения для других метрик и типов недифференцируемых моделей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Raj, A., Sharma, V., & Shanu, A.K. (2022). Comparative analysis of security and privacy technique for federated learning in IOT based devices. *2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 1-5. DOI: https://doi.org/doi:10.1109/ICCAKM54721.2022.9990152

2 Reddy, G.P., & Pavan Kumar, Y.V. (2023). A Beginner's guide to federated learning. *2023 Intelligent Methods, Systems, and Applications (IMSA)*, 557-562. DOI: https://doi.org/doi:10.1109/IMSA58542.2023.10217383

3 Panigrahi, M., Bharti, S., & Sharma, A. (2023). Federated learning for beginners: types, simulation environments and open challenges. *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3)*, 1-6. DOI: https://doi.org/doi:10.1109/IC2E357697.2023.10262769

4 Pang, S., Peng, Y., Ban, T., Inoue, D., & Sarrafzadeh, A. (2015). A federated network online network traffics analysis engine for cybersecurity. 2015 International Joint Conference on Neural Networks (IJCNN), 1-8. DOI: https://doi.org/doi:10.1109/IJCNN.2015.7280563

5 Mosharraf, M., & Taghiyareh, F. (2017). Federated search engine for open educational linked data. URL: https://www.researchgate.net/publication/318786955_Federated_search_engine_for_open_educational_linked_data

6 Smith, V., Chiang, C., Sanjabi, M., & Talwalkar, A. (2017). Federated multi-task learning. neural information processing systems. URL: https://api.semanticscholar.org/CorpusID:3586416

7 Zhou, W., Li, Y., Chen, S., & Ding, B. (2018). Real-time data processing architecture for multi-robots based on differential federated learning. 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 462-471. DOI: https://doi.org/doi:10.1109/SmartWorld.2018.00106

8 Ekellem, E.A. (2023). Secure Federations: Addressing the security challenges in federated learning and privacy-preserving AI. *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 1-6. DOI: https://doi.org/doi:10.1109/ISAS60782.2023.10391351

9 Aledhari, M., Razzak, R., Parizi, R.M., & Saeed, F. (2020). Federated learning: a survey on enabling technologies, protocols and applications. IEEE Access, 8, 140699-140725. DOI: https://doi.org/doi:10.1109/ACCESS.2020.3013541

10 Supriya, Y., & Gadekallu, T.R. (2023). A survey on soft computing techniques for federated learning- applications, challenges and future directions. ACM Journal of Data and Information Quality, 15, 1 - 28. DOI: https://doi.org/10.1145/3575810

11 Li, T., Sahu, A., Talwalkar, A., & Smith, V. (2019). Federated learning: challenges, methods and future directions. *IEEE Signal Processing Magazine, 37*, 50-60. DOI: https://doi.org/doi:10.1109/MSP.2020.2975749
12 Tran, V., Pham, H., & Wong, K. (2023). Personalized privacy-preserving framework for cross-silo federated learning. *IEEE Transactions on Emerging Topics in Computing, 12*, 1014-1024. DOI: https://doi.org/doi:10.1109/TETC.2024.3356068
13 Dinh C. Nguyen, et al (2021). Federated Learning for internet of things: a comprehensive survey. IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 23, NO. 3. DOI: https://doi.org/doi:10.1109/COMST.2021.3075439
14 McMahan, H.B., Moore, E., Ramage, D., Hampson, S., & Arcas, B.A. (2016). Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics.* URL: https://api.semanticscholar.org/CorpusID:14955348
15 Kang, J., Xiong, Z., et al (2019). Reliable Federated Learning for mobile networks. *IEEE Wireless Communications, 27*, 72-80. DOI: https://doi.org/doi:10.1109/MWC.001.1900119
16 Kairouz, P., et al (2019). Advances and open problems in Federated Learning. *Found. Trends Mach. Learn., 14*, 1-210. DOI: https://doi.org/10.48550/arXiv.1912.04977
17 Dinh, T.Q., Tran, T., & Le, T. (2021). Communication cost reduction using sparse ternary compression and encoding for FedAvg. 2021 International Conference on Information and Communication Technology Convergence (ICTC), 351-356. DOI: https://doi.org/doi:10.1109/ICTC52510.2021.9620887
18 Kairouz, P., *et al* (2019). Advances and open problems in federated learning. *Found. Trends Mach. Learn., 14*, 1-210. URL: https://api.semanticscholar.org/CorpusID:209202606
19 Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. *ArXiv, abs/1806.00582*. URL: https://api.semanticscholar.org/CorpusID:46936175
20 Bonawitz, Keith et al. Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017). URL: https://api.semanticscholar.org/CorpusID:3833774
21 Mao, Y., Zhang, J., & Letaief, K.B. (2016). Dynamic computation offloading for mobile-edge computing with energy harvesting devices. *IEEE Journal on Selected Areas in Communications, 34*, 3590-3605. https://doi.org/doi:10.1109/JSAC.2016.2611964
22 Chen, Z., Lv, N., Liu, P., Fang, Y., Chen, K., & Pan, W. (2020). Intrusion detection for wireless edge networks based on Federated Learning. *IEEE Access, 8*, 217463-217472. URL: https://api.semanticscholar.org/CorpusID:228098118
23 Bonawitz, K., Eichner, et al (2019). Towards federated learning at scale: system design. *ArXiv, abs/1902.01046*. URL: https://api.semanticscholar.org/CorpusID:59599820

24 Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H.V., & Cui, S. (2019). A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications, 20*, 269-283. DOI: https://doi.org/doi:10.1109/TWC.2020.3024629
25 Bonawitz, K., Eichner, H., et al (2019). Towards Federated Learning at scale: system design. *ArXiv, abs/1902.01046*. URL: https://api.semanticscholar.org/CorpusID:59599820
26 Konecný, J., McMahan, H.B., et al (2016). Federated learning: strategies for improving communication efficiency. *ArXiv, abs/1610.05492*. URL: https://api.semanticscholar.org/CorpusID:14999259
27 Konecný, J., McMahan, H.B., Ramage, D., & Richtárik, P. (2016). Federated optimization: distributed machine learning for on-device intelligence. *ArXiv, abs/1610.02527*. URL: https://api.semanticscholar.org/CorpusID:2549272
28 Khan, L.U., Saad, W., Han, Z., Hossain, E., & Hong, C.S. (2020). Federated learning for internet of things: recent advances, taxonomy and open challenges. *IEEE Communications Surveys & Tutorials, 23*, 1759-1799. DOI: https://doi.org/doi:10.1109/COMST.2021.3090430
29 Niknam, S., Dhillon, H.S., & Reed, J.H. (2019). Federated learning for wireless communications: motivation, opportunities and challenges. *IEEE Communications Magazine, 58*, 46-51. DOI: https://doi.org/doi:10.1109/MCOM.001.1900461
30 Kale, D.R., Mane, T., Buchade, A., Patel, P., Wadhwa, L.K., & Pawar, R.G. (2024). Federated Learning for privacy-preserving data mining. *2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA)*, 1-6. DOI: https://doi.org/doi:10.1109/ICISAA62385.2024.10828741
31 Sheller, M.J., Edwards, et al (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports, 10*. URL: https://api.semanticscholar.org/CorpusID:220812287
32 Nguyen, D.C., Pham, V.Q., et al (2021). Federated Learning for smart healthcare: a survey. *ACM Computing Surveys (CSUR), 55*, 1-37. URL: https://api.semanticscholar.org/CorpusID:244270523
33 Liu, Y., Kumar, N., Xiong, Z., Lim, W.Y., Kang, J., & Niyato, D.T. (2020). Communication-efficient federated learning for anomaly detection in industrial internet of things. *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 1-6. URL: https://api.semanticscholar.org/CorpusID:248604676
34 Canetti, R., Feige, U., Goldreich, O., & Naor, M. (1996). Adaptively secure multi-party computation. *Symposium on the Theory of Computing*. URL: https://api.semanticscholar.org/CorpusID:2840228
35 Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Gener. Comput. Syst., 115*, 619-640. URL: https://api.semanticscholar.org/CorpusID:225140677
36 Dhumale, S., Ameria, D., & Shukla, S. (2023). Blockchain-based personalized federated learning leveraging the SMPC protocol. *2023 IEEE International*

Carnahan Conference on Security Technology (ICCST), 1-6. DOI: https://doi.org/doi:10.1109/ICTC52510.2021.9620887
37 Chen, J., Chen, M., Zeng, G., & Weng, J. (2021). BDFL: A byzantine-fault-tolerance decentralized federated learning method for autonomous vehicle. *IEEE Transactions on Vehicular Technology, 70*, 8639-8652. DOI: https://doi.org/doi:10.1109/TVT.2021.3102121
38 Liu, Z., Sun, H., Song, J., Zhang, B., Yan, Y., Qiu, B., Jiang, L., & Li, J. (2023). Vertical federated learning architecture for power company and financial company and electricity pricing model considering user credit evaluation. *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 820-826. DOI: https://doi.org/doi:10.1109/ICCECE58074.2023.10135197
39 Abidin, N.Z., & Ritahani Ismail, A. (2022). Federated deep learning for automated detection of diabetic retinopathy. *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)*, 1-5. DOI: https://doi.org/doi:10.1109/ICCED56140.2022.10010636
40 Ullah, F., Srivastava, G., Xiao, H., Ullah, S., Lin, J.C., & Zhao, Y. (2023). A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging. *IEEE Journal of Biomedical and Health Informatics, 28*, 3293-3304. DOI: https://doi.org/doi:10.1109/JBHI.2023.3282955
41 Pereira, K., Parikh, A., Kumar, P.P., & Devadkar, K. (2023). Healthcare diagnostics service using Federated Learning. *2023 International Conference for Advancement in Technology (ICONAT)*, 1-6. DOI: https://doi.org/doi:10.1109/ICONAT57137.2023.10080053
42 R, A., Renuka D, K., S, O., & R, S. (2023). Secured data sharing of medical images for disease diagnosis using deep learning models and federated learning framework. *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 499-504. DOI: https://doi.org/doi:10.1109/ICISCoIS56541.2023.10100542
43 Patil, A., Choudhar, A., Shah, D., Abraham, J., & Bochare, A.P. (2024). Analyzing poisoning attacks on non-IID federated learning systems for credit scoring. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-7. DOI: https://doi.org/doi:10.1109/ICCCNT61001.2024.10724119
44 Mukisa, K.J., Ahakonye, L.A., Kim, D., & Lee, J.M. (2024). Trimmed averaging for efficient federated learning in the internet of things. *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, 322-326. DOI: https://doi.org/doi:10.1109/ICTC62082.2024.10827800
45 Wu, H., Xu, Z., Ding, Y., Shi, W., Zhang, A., & Liu, Y. (2024). Graph convolutional network-based federated learning method for distribution system topology identification. *2024 China International Conference on Electricity Distribution (CICED)*, 724-728. DOI: https://doi.org/doi:10.1109/CICED63421.2024.10754175

46 Ojha, A.C., Kumar Yadav, D., & B, A. (2023). Federated learning paradigms in network security for distributed systems. *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 1-5. DOI: https://doi.org/doi:10.1109/ICTBIG59752.2023.10456162
47 Kairouz, P. *et al* (2019). Advances and open problems in federated learning. *Found. Trends Mach. Learn., 14*, 1-210. URL: https://doi.org/10.48550/arXiv.1912.04977
48 Saha, S., & Ahmad, T. (2020). Federated transfer learning: concept and applications. *Intelligenza Artificiale, 15*, 35-44. URL: https://api.semanticscholar.org/CorpusID:225103353
49 Khan, L.U., Saad, W., Han, Z., Hossain, E., & Hong, C.S. (2020). Federated learning for internet of things: recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials, 23*, 1759-1799. DOI: https://doi.org/doi:10.1109/COMST.2021.3090430
50 Kang, J., Xiong, Z., Niyato, D.T., Zou, Y., Zhang, Y., & Guizani, M. (2019). Reliable federated learning for mobile networks. *IEEE Wireless Communications, 27*, 72-80. DOI: https://doi.org/doi:10.1109/MWC.001.1900119
51 Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2019). On the convergence of FedAvg on non-IID data. *ArXiv, abs/1907.02189*. URL: https://api.semanticscholar.org/CorpusID:195798643
52 Wang, K., Ding, Z., So, D.K., & Ding, Z. (2024). Energy efficient federated learning with age-weighted FedSGD. *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 457-462. DOI: https://doi.org/doi:10.1109/ICCWorkshops59551.2024.10615715
53 Zhao, Y., Zhao, T., Xiang, P., Li, Q., & Chen, Z. (2023). Multi-task federated learning medical analysis algorithm integrated into adapter. *2023 IEEE 8th International Conference on Big Data Analytics (ICBDA)*, 24-30. DOI: https://doi.org/10.1109/ICBDA57405.2023.10104867
54 Sad, C., Retsinas, G., Soudris, D., Siozios, K., & Masouros, D. (2025). Towards asynchronous peer-to-peer federated learning for heterogeneous systems. *Proceedings of the 5th Workshop on Machine Learning and Systems*. DOI: https://doi.org/10.1145/3721146.3721952
55 Jiang, J., Shi, S., Li, Y., Wang, F., & Zhang, Y. (2023). Hierarchical semi-asynchronous federated learning based on over-the-air computation. *2023 IEEE 15th International Conference on Advanced Infocomm Technology (ICAIT)*, 73-78. DOI: https://doi.org/doi:10.1109/ICAIT59485.2023.10367298
56 Wei, T., Wang, Y., & Li, W. (2022). The deep flow inspection framework based on horizontal Federated Learning. *2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 1-4. DOI: https://doi.org/doi:10.23919/APNOMS56106.2022.9919969
57 Li, L., Fan, Y., Tse, M., & Lin, K.-Y. (2020). A review of applications in federated learning. *Computers & Industrial Engineering*, *149*, 106854. DOI: https://doi.org/10.1016/j.cie.2020.106854

58 Gao, J., Ning, Z., Cui, M., & Xing, S. (2024). An efficient and security Federated Learning for data heterogeneity. *2024 4th International Conference on Information Communication and Software Engineering (ICICSE)*, 1-5. DOI: https://doi.org/doi:10.1109/ICICSE61805.2024.10625682
59. Pingulkar, S., & Pawade, D.Y. (2024). Federated learning architectures for credit risk assessment: a comparative analysis of vertical, horizontal, and transfer learning approaches. *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, 1-7. DOI: https://doi.org/doi:10.1109/ICBDS61829.2024.10837430
60 Liu, D., Bai, L., Yu, T., & Zhang, A. (2022). Towards method of horizontal federated learning: a survey. *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, 259-266. DOI: https://doi.org/doi:10.1109/BigDIA56350.2022.9874186
61 Serpanos, D., & Xenos, G. (2023). Vertical federated learning in malware detection for smart cities. *2023 IEEE International Smart Cities Conference (ISC2)*, 1-5. DOI: https://doi.org/doi:10.1109/ISC257844.2023.10293429
62 Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y., & Yang, Q. (2022). Vertical federated learning: concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering, 36*, 3615-3634. DOI: https://doi.org/10.1109/TKDE.2024.3352628
63 Khan, A., Thij, M.T., Thuijsman, F., & Wilbik, A. (2023). Using the nucleolus for incentive allocation in vertical federated learning. *2024 2nd International Conference on Federated Learning Technologies and Applications (FLTA)*, 224-231 DOI: https://doi.org/doi:10.1109/FLTA63145.2024.10839765
64 Wang, L., Liu, L., Lu, Y., Zhang, C., Zheng, Y., & Xu, J. (2023). Hierarchical federated learning for heterogeneous features and distributed data in IoT networks. *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, 1-6. DOI: https://doi.org/doi:10.1109/ICCC57788.2023.10233409
65 Bektemyssova, G. U., Bakirova, G. S., et al (2024). Analysis of the relevance and prospects of federated learning application. *Bulletin of the National Academy of Sciences of the Republic of Kazakhstan, 2*(92), 56-65. DOI: https://doi.org/doi:10.47533/2024.1606-146X.26
66 Annunziata, D., Canzaniello, M., Savoia, M., Cuomo, S., & Piccialli, F. (2024). Benchmarking federated learning on high-performance computing: aggregation methods and their impact. *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 207-214. DOI: https://doi.org/doi:10.1109/PDP62718.2024.00036
67 Simos, M., Bouzinis, P.S., Diamantoulakis, P.D., Sarigiannidis, P.G., & Karagiannidis, G.K. (2022). Hierarchical federated learning for the next generation IoT. *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 198-203. https://doi.org/doi:10.1109/WiMob55322.2022.9941355
68 Delehouzée, M., Lessage, X., Reginster, T., & Mahmoudi, S. (2024). Performance analysis of aggregation algorithms in cross-silo federated learning for

non-IID data. *2024 4th International Conference on Embedded & Distributed Systems (EDiS)*, 74-79. DOI: https://doi.org/doi:10.1109/EDiS63605.2024.10783224
69 Majeed, U., Hassan, S.S., & Hong, C.S. (2021). Cross-silo model-based secure federated transfer learning for flow-based traffic classification. *2021 International Conference on Information Networking (ICOIN)*, 588-593. DOI: https://doi.org/doi:10.1109/ICOIN50884.2021.9333905
70 Yu, S., Nguyen, P.M., Anwar, A., & Jannesari, A. (2021). Heterogeneous federated learning using dynamic model pruning and adaptive gradient. *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 322-330. DOI: https://doi.org/10.48550/arXiv.2106.06921
71 Pingulkar, S., & Pawade, D.Y. (2024). Federated learning architectures for credit risk assessment: a comparative analysis of vertical, horizontal, and transfer learning approaches. *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, 1-7. DOI: https://doi.org/doi:10.1109/ICBDS61829.2024.10837430
72 Wu, R., Mitra, S., Chen, X., & Rao, A. (2023). Decentralized personalized online federated learning. *2023 IEEE International Conference on Big Data (BigData)*, 1873-1882. DOI: https://doi.org/10.48550/arXiv.2311.04817
73 Tyagi, S., Rajput, I.S., & Pandey, R. (2023). Federated learning: applications, security hazards and defense measures. *2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT)*, 477-482. DOI: https://doi.org/doi:10.1109/DICCT56244.2023.10110075
74 Vajjhula, R.V., Alawadi, S., Goswami, P., & Buravelli, S.K. (2025). A comparative study of federated learning Methods for human activities recognition in healthcare. *2025 7th International Conference on Blockchain Computing and Applications (BCCA)*, 729-736. DOI: https://doi.org/10.1109/BCCA66705.2025.11229651
75 Tummala, V.M., Hazra, A., Kalita, A., & Mohan, G. (2024). Cluster based pseudo hierarchical decentralized federated learning in UAV networks. *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 1-5. DOI: https://doi.org/doi:10.1109/VTC2024-Fall63153.2024.10757781
76 Gupta, D., Dhanda, N., & Gupta, K.K. (2025). A comparative study of various LSTM models for stock market time series classification. *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)*, 1-6. DOI: https://doi.org/doi: 10.1109/ICCMC65190.2025.11140645
77 Николенко, С. И., Кадурин, А. А., & Архангельская, Е. В. (2018). *Глубокое обучение. Погружение в мир нейронных сетей*. Санкт-Петербург: Питер. URL:https://habr.com/ru/companies/piter/articles/346358/
78 Kumar, C.C., & Parthipan, V. (2024). Performance analysis of predicting LIC stock price using Lasso regression compared with Random Forest Regression. *2024 Second International Conference Computational and Characterization Techniques*

in Engineering & Sciences (IC3TES), 1-5. DOI: https://doi.org/10.1109/IC3TES62412.2024.10877466
79 Nourmohammadi, R., Sabramooz, M.R., Zhang, K., & Talhi, C. (2024). BlockFed: a novel federated learning framework based on hierarchical aggregation. *2024 6th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, 1-4. DOI: https://doi.org/10.1145/3701717.3730545
80 Tyagi, S., Rajput, I.S., & Pandey, R. (2023). Federated learning: applications, security hazards and defense measures. *2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT)*, 477-482. https://doi.org/doi:10.1109/DICCT56244.2023.10110075
81 Fu, T., Luo, J., Chen, K., Du, N., Gong, X., & Wang, J. (2023). CatBoost ridge regression based sailboat price forecasting. *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, 385-389. DOI: https://doi.org/10.1109/ICSECE58870.2023.10263515
82 Gorde, P.S., & Nilesh Borkar, S. (2024). Comparative analysis of linear regression, random forest regressor and LSTM for stock price prediction. *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1-5. DOI: https://doi.org/doi:10.1109/ICCUBEA61740.2024.10775094
83 Sadaf, K. (2023). Phishing website detection using XGBoost and Catboost classifiers. *2023 International Conference on Smart Computing and Application (ICSCA)*, 1-6. DOI: https://doi.org/doi:10.1109/ICSCA57840.2023.10087829
84 Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., & Jararweh, Y. (2022). Federated learning review: fundamentals, enabling technologies and future applications. *Inf. Process. Manag., 59*, 103061. DOI: https://doi.org/10.1016/j.ipm.2022.103061
85 Banabilah, S., Aloqaily, M., et al (2022). Federated learning review: fundamentals, enabling technologies and future applications. *Inf. Process. Manag., 59*, 103061. DOI: https://doi.org/10.1016/j.ipm.2022.103061
86 Reddi, S.J., et al (2020). Adaptive federated optimization. *ArXiv, abs/2003.00295*. DOI: https://doi.org/10.48550/arXiv.2003.00295
87 Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): european regulation that has a global impact. *International Journal of Market Research*, *59*(6), 703-705. URL: https://gdpr-info.eu/
88 Su, L., Xu, J., & Yang, P. (2021). A non-parametric view of FedAvg and FedProx: beyond stationary points. *J. Mach. Learn. Res., 24*, 203:1-203:48. URL: https://jmlr.org/papers/v24/22-0153.html
89 Herlambang, S.W., Dewanta, F., & Purwanto, Y. (2025). Federated learning approaches for iot intrusion detection based on FedAvg and FedProx on IID and non-IID data. *2025 International Conference on Information and Communication Technology (ICoICT)*, 1-6. DOI: https://doi.org/10.1109/ICoICT66265.2025.11192987

90 Sahu, A., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., & Smith, V. (2018). Federated optimization in heterogeneous networks. *arXiv: Learning*. DOI: https://api.semanticscholar.org/CorpusID:59316566
91 Tang, Z., & Chang, T. (2024). FedLion: faster adaptive federated optimization with fewer communication. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 13316-13320. https://doi.org/10.48550/arXiv.2402.09941
92 Bonawitz, K., Eichner, H., et al (2019). Towards federated learning at scale: system design. *ArXiv, abs/1902.01046*. URL: https://api.semanticscholar.org/CorpusID:59599820
93 H. Reguieg, M. E. Hanjri, M. E. Kamili and A. Kobbane. (2023). A comparative evaluation of FedAvg and Per-FedAvg algorithms for dirichlet distributed heterogeneous data, 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Istanbul, Turkiye, 2023, pp. 1-6, DOI: https://doi.org/10.48550/arXiv.2309.01275
94 Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnovic, M. (2016). QSGD: communication-efficient SGD via gradient quantization and encoding. *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:263894534
95 Bernstein, J., Wang, Y., Azizzadenesheli, K., & Anandkumar, A. (2018). signSGD: compressed optimisation for non-convex problems. *International Conference on Machine Learning*. URL: https://api.semanticscholar.org/CorpusID:7763588
96 Su, L., Zhou, R., Wang, N., Chen, J., & Li, Z. (2024). Low-latency hierarchical federated learning in wireless edge networks. *IEEE Internet of Things Journal, 11*, 6943-6960. DOI: https://doi.org/doi:10.1109/JIOT.2023.3314743
97 Reddi, S.J., Charles, Z.B., et al (2020). Adaptive federated optimization. *ArXiv, abs/2003.00295*. URL: https://api.semanticscholar.org/CorpusID:211678094
98 Sahu, A., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., & Smith, V. (2018). Federated optimization in heterogeneous networks. *arXiv: Learning*. DOI: https://doi.org/10.48550/arXiv.1812.06127
99 Ефремов, М.А., Холод, И. (2022). Разработка архитектуры универсального фреймворка федеративного обучения. Программные продукты и системы, 35 (2), 263-272. DOI: https://doi.org/doi:10.15827/0236-235X.138.263-272
100 Yin, T., Li, L., Lin, W., Liang, W., Li, X., & Han, Z. (2023). FedSCS: client selection for federated learning under system heterogeneity and client fairness with a stackelberg game approach. *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 373-378. DOI: https://doi.org/doi:10.1109/ICCWorkshops57953.2023.10283698
101 Gu Y. (2024). A comparative analysis study of stock prediction based on random forest and decision tree. *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, Marseille, France, 2024, 96-100, DOI: https://doi.org/doi:10.1109/ICEDCS64328.2024.00022

102 Sabbah, I., Sabbah, H., et al (2012). Burnout among lebanese nurses: psychometric properties of the maslach burnout inventory-human services survey (MBI-HSS). *Health, 4*, 644-652. DOI: https://doi.org/doi:10.4236/health.2012.49101
103 Kokkinos, C.M. (2006). Factor structure and psychometric properties of the maslach burnout inventory educators survey among elementary and secondary school teachers in Cyprus. *Stress and Health, 22*, 25-33. DOI: https://doi.org/10.1002/smi.1079
104 Loera, B., Converso, D., & Viotti, S. (2014). Evaluating the psychometric properties of the maslach burnout inventory human services survey (MBI HSS) among italian nurses: how many factors must a researcher consider? 9(12). DOI:https://doi.org/10.1371/journal.pone.0114987
105 Puranitee, P., Saetang, S., Sumrithe, et al (2019). Exploring burnout and depression of Thai medical students: the psychometric properties of the maslach burnout inventory. International Journal of Medical Education, 10, 223–229. https://doi.org/10.5116/ijme.5dc6.8228

ПРИЛОЖЕНИЕ А

Исходный код

Код на стороне сервера

```
# Установка Flask
!pip install flask
# Установка Node.js и npm (только если они еще не установлены)
!apt update
!apt install -y nodejs npm
# Установка LocalTunnel
!npm install -g localtunnel
from flask import Flask, request, jsonify
import json
import os
import threading
import subprocess
import time
app = Flask(__name__)
# Путь к файлу для сохранения обновлений
RECEIVED_UPDATES_PATH = "received_updates.json"
# Функция для запуска LocalTunnel и получения публичного URL
def start_localtunnel():
process = subprocess.Popen(["lt", "--port", "5010"], stdout=subprocess.PIPE,
stderr=subprocess.PIPE, text=True)
for line in process.stdout:
if "https://" in line:
public_url = line.strip().split(' ')[-1]
print(f"Туннель открыт по URL: {public_url}")
break
# Главный маршрут для тестирования сервера
@app.route('/update')
def home():
  print("Запрос на главный маршрут")
  return "Hello, Flask is running through LocalTunnel!"
# Маршрут для приема обновлений модели
@app.route('/update', methods=['POST'])
def update_model():
data = request.get_json()
if data is None:
print("Ошибка: Пустой JSON получен от клиента.")
return jsonify({"error": "Invalid JSON"}), 400
client_id = data.get("client_id", "unknown_client")
print(f"Получено обновление модели от клиента: {client_id}")
# Сохранение обновлений от клиента в файл
if os.path.exists(RECEIVED_UPDATES_PATH):
with open(RECEIVED_UPDATES_PATH, "r") as f:
all_updates = json.load(f)
else:
all_updates = []
all_updates.append(data)
```

```
with open(RECEIVED_UPDATES_PATH, "w") as f:
json.dump(all_updates, f)
print(f"Обновление модели от клиента {client_id} успешно сохранено в
{RECEIVED_UPDATES_PATH}")
return jsonify({"message": f"Updates received from {client_id} and saved!"}), 200
# Функция для запуска Flask в отдельном потоке
def run_flask():
app.run(port=5010)
if __name__ == "__main__":
# Запуск LocalTunnel в отдельном потоке
tunnel_thread = threading.Thread(target=start_localtunnel)
tunnel_thread.daemon = True
tunnel_thread.start()
# Запуск Flask в отдельном потоке
flask_thread = threading.Thread(target=run_flask)
flask_thread.daemon = True
flask_thread.start()
# Основной код, выполняемый после запуска Flask
print("Flask сервер запущен через LocalTunnel. Выполняем основной код")
time.sleep(5)
print("Основной код завершен.")
```

```
Flask сервер запущен через LocalTunnel. Выполняем основной код
* Serving Flask app '__main__'
* Debug mode: off
INFO:werkzeug:WARNING: This is a development server. Do not use it in a production
deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5010
INFO:werkzeug:Press CTRL+C to quit
Туннель открыт по URL: https://yellow-rice-stand.loca.lt
Основной код завершен.
```

Открывает сервер

In []:

```
import os
# Удаляем файлы
os.remove('/content/global_model.json')
os.remove('/content/model0_updates.json')
os.remove('/content/model1_updates.json')
os.remove('/content/model2_updates.json')
os.remove('/content/received_updates.json')
print("Файл model0_updates, model1_updates, model2_updates, global_model,
received_updates.json успешно удалены.")
```

```
Файл model0_updates, model1_updates, model2_updates, global_model, received_updates.json
успешно удалены.
```

Код доя приема нескольких моделей

In [ ]:

```
from flask import Flask, request, jsonify
import json
import os
import threading
import subprocess
import time
```

```
app = Flask(__name__)
# Путь для сохранения файлов с обновлениями моделей
UPDATE_FILES = {
"Model Psycho-emotional exhaustion": "model0_updates.json",
"Model Depersonalization": "model1_updates.json",
"Model Reduction of personal achievements": "model2_updates.json"
# Функция для запуска LocalTunnel и получения публичного URL
def start_localtunnel():
process = subprocess.Popen(["lt", "--port", "5021"], stdout=subprocess.PIPE,
stderr=subprocess.PIPE, text=True)
for line in process.stdout:
if "https://" in line:
public_url = line.strip().split(' ')[-1]
print(f"Туннель открыт по URL: {public_url}")
break
# Главный маршрут для тестирования сервера
@app.route('/')
def home():
print("Запрос на главный маршрут")
return "Hello, Flask is running through LocalTunnel!"
# Маршрут для приема обновлений модели
@app.route('/update', methods=['POST'])
def update_model():
data = request.get_json()
if data is None:
print("Ошибка: Пустой JSON получен от клиента.")
return jsonify({"error": "Invalid JSON"}), 400
model_name = data.get("model", "unknown_model")
filename = UPDATE_FILES.get(model_name)
if filename is None:
print(f"Неизвестная модель: {model_name}. Обновления не сохранены.")
return jsonify({"error": "Unknown model"}), 400
print(f"Получено обновление для модели '{model_name}'")
# Чтение текущих обновлений из файла
if os.path.exists(filename):
with open(filename, "r") as f:
all_updates = json.load(f)
else:
all_updates = []
# Добавление новых данных и сохранение в файл
all_updates.append(data)
try:
with open(filename, "w") as f:
json.dump(all_updates, f)
print(f"Обновление для модели '{model_name}' успешно сохранено в {filename}")
return jsonify({"message": f"Updates for '{model_name}' received and saved!"}), 200
except IOError as e:
print(f"Ошибка записи в файл {filename}: {e}")
return jsonify({"error": "Failed to save updates"}), 500
# Функция для запуска Flask в отдельном потоке
def run_flask():
app.run(port=5025)
```

```
if __name__ == "__main__":
# Запуск LocalTunnel в отдельном потоке
tunnel_thread = threading.Thread(target=start_localtunnel)
tunnel_thread.daemon = True
tunnel_thread.start()
# Запуск Flask в отдельном потоке
flask_thread = threading.Thread(target=run_flask)
flask_thread.daemon = True
flask_thread.start()
# Основной код, выполняемый после запуска Flask
print("Flask сервер запущен через LocalTunnel. Выполняем основной код...")
time.sleep(5)
print("Основной код завершен.")
Flask сервер запущен через LocalTunnel. Выполняем основной код...
 * Serving Flask app '__main__'
 * Debug mode: off
INFO:werkzeug:WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5025
INFO:werkzeug:Press CTRL+C to quit
Туннель открыт по URL: https://quick-items-chew.loca.lt
Основной код завершен.
Вывод содежимого model0_updates.json, model1_updates.json, и model2_updates.json
In [ ]:
import json
# Список файлов для обработки
file_paths = ["model0_updates.json", "model1_updates.json", "model2_updates.json"]
# Чтение и вывод данных для каждого файла
for file_path in file_paths:
  print(f"\nЧтение данных из файла {file_path} ")
try:
# Чтение данных из файла
with open(file_path, "r") as f:
updates = json.load(f)
# Проверка, является ли содержимое массивом записей
if isinstance(updates, list):
for index, update in enumerate(updates):
print(f"\nСодержимое записи {index+1} в {file_path}:")
model_name = update.get("model", "Неизвестная модель")
print(f"Модель: {model_name}")
# Выводим важности признаков
model_params = update.get("model_params", {})
feature_importances = model_params.get("feature_importances", "Нет данных")
print("Важности признаков:", feature_importances)
# Выводим метрики
metrics = update.get("metrics", {})
print("Метрики:")
for metric, value in metrics.items():
print(f" {metric.upper()}: {value}")
else:
print(f"Ошибка: Ожидался массив записей в {file_path}, но получен другой формат.")
```

except (FileNotFoundError, json.JSONDecodeError) as e:
print(f"Ошибка при обработке файла {file_path}: {e}")
Чтение данных из файла model0_updates.json...
Содержимое записи 1 в model0_updates.json:
Модель: Model Psycho-emotional exhaustion
Важности признаков: [0.09178429567706442, 0.023627553510125264, .07992842854717708,
0.021190004072464, 0.0801755819460731, 0.0031318930763287282, 0.006668890414854371,
0.008210305927271207, 0.00379559928639662, 0.005795472959355808, 0.006336267350631,
0.002224100614238205, 0.005519564181658134, 0.011391965805351179,
0.004719626032935, 0.0085627595061539, 0.003915251001708355, 0.011115093886978395,
0.002339809036179903 2, 0.006553656099343803, 0.010055758599015436, 0.0045960892739,
0.005287892782532482, 0.0031882696369227763, 0.005648929542470493,
0.00222844800543838, 0.004826647914383728, 0.006860549825683, 0.006212800387773212,
0.005525839173395509, 0.0021782640281968126, 0.0022553343007405834,
0.005534687550710736 5, 0.002806102194298814, 0.0058990325139, 0.008013583844280442,
0.0032073880857734004, 0.0011829083639454758, 0.0009977081569643742,
0.006039857395390127, 0.00408434551163553, 0.001715846390664, 0.005236950405167964,
0.0030943029816562, 0.002386867956759, 0.002181429485316705, 0.0019209057241942312,
0.002465846089621358, 0.0025450145778036187, 0.00315152833149108,
0.0010506131237563619, 0.0006604511394597, 0.005421807532924,
0.0016262029323379414, 0.006422897747099592, 0.00185188533541852,
0.00251326544700154, 0.0049620956910475, 0.0016848387837486, 0.0037198360558487785,
0.00034977788189122596, 0.003504012463397, 0.004203497967282248,
0.0028993486836704, 0.0066772651436822, 0.001974702548042013, 0.005369179547377387,
0.007129242525342203, 0.0027616821384567063, 0.0043412083480005945,
0.00220690059033, 0.007007956929728006, 0.00091722978289393, 0.0002529598390440906,
0.0005221831436975996, 0.003858299161151416, 0.002186001722186556,
0.00131460677166757, 0.0014541166081096437, 0.0028018262115357445,
0.0011171592168854353, 0.0020215003227621034, 0.0030803840620888512,
0.006575664546734358, 0.0029499184659267674, 0.005919648563618224,
0.0018951418083437616, 0.0016256326062463766, 0.003083835468040148,
0.0027953818079130112, 0.004262712838607021, 0.002742323201463146,
0.007694385618757738, 0.00497867715806649, 0.007285834530447965,
0.003585658390862064, 0.005580321420761033, 0.0033743883779794575,
0.0036881888833684524, 0.0064866555345492885, 0.001749310670333388,
0.00238008597116879, 0.002753741793542859, 0.004558666657378093,
0.018264352995323554, 0.007034783511945225, 0.004378044938351417,
0.004133092098649946, 0.005718802660122449, 0.001855335265959257,
0.003962451159868126, 0.003287265035296021, 0.0024448341059000395,
0.0035927189615375748, 0.0030580372939235808, 0.003936806856249645,
0.00726024430108296, 0.007937838755086276, 0.006755367574060914,
0.0031220472583828574, 0.002755203328337259, 0.00963670190367486,
0.00536197152951697, 0.0034860846278129726, 0.0007323947332602451,
0.005616447660945477, 0.005509025173203063, 0.0044212905082577185,
0.0063530092241028905, 0.005355608747398939, 0.005430210857469536,
0.005762513731881922, 0.00770281944552723, 0.0035467356669713662,
0.0071051417155456095, 0.0032920575749000518, 0.007531897914625263,
0.008557570998173943, 0.0022585220165258417, 0.007516545309784434,
0.002662529096313971, 0.0021335462831680 22, 0.005911448344131375,
0.00526121503781069, 0.011303925770924988, 0.007212531169615507,
0.009059770994445918, 0.003169393948297719, 0.0023335620864195373,
0.0023927796056730983, 0.0022742947105256904, 0.005946128965519041,

0.0010094655570735314, 0.0048249738927195205, 0.000360488777043646,
0.004267139597475676, 0.00757939466884326, 0.004072610851542213,
0.004153352580055614, 0.0025271447884859842, 0.0036407981410245625,
0.0038666446837695885, 0.0047030425459468304, 0.0007878302187226496,
0.002665928697997914, 0.00148674159679158, 0.005554611788744628,
0.002017910380711376]
Метрики:
ACCURACY: 0.9578378378378378
PRECISION: 0.9540918163672655
RECALL: 0.9676113360323887
F1_SCORE: 0.9608040201005025
MSE: 0.03855324324324324
R2: 0.8450683433217168
Содержимое записи 2 в model0_updates.json:
Модель: Model Psycho-emotional exhaustion
Важности признаков: [0.3427282796114314, 0.032653859454048924, 0.25986959847120533,
0.02304235980397729, 0.314957497719371, 0.0, 1.1655964946778063e-09,
7.49327981786134e-05, 4.519639923934861e-06, 0.00017528838025611184,
2.8179637193607566e-06, 1.4418765363938804e-05, 1.6218763342572848e-06,
4.1196665372999485e-06, 5.265513210965696e-06, 5.454002603552338e-06,
3.4124515453380193e-06, 2.4418206054889266e-06, 2.0378719418315926e-06,
0.0003074909180429705, 1.918074618288593e-06, 3.8569236693885176e-06,
0.0005690551234377861, 6.223654328292512e-06, 2.647725998774107e-06,
2.1182599115822465e-05, 0.0007333109991040713, 0.00018295795619663178,
2.532975604273345e-05, 3.1229676271323154e-06, 1.3931639201518034e-06,
1.323231746016003e-06, 6.675924826533557e-08, 4.767500309143662e-05,
1.559389288367005e-05, 6.777183548164929e-05, 9.991231915025619e-08,
4.241331637455188e-08, 0.00017568040245887636, 4.1069325717599643e-05,
1.505868840859778e-06, 0.002006472669063346, 0.00010267602824347882,
0.000459304156027388, 1.7628991920545486e-06, 7.351224375916193e-06,
4.6030113466278713e-07, 3.760016520490823e-07, 2.051194947169487e-05,
3.584415616046351e-07, 3.3874443401020377e-06, 1.096263619664631e-05,
4.296362636719961e-07, 9.022679635729721e-06, 4.239975642587534e-06,
2.0006361154330418e-05, 2.6919000083164894e-05, 1.0841558369874685e-06,
2.9572722561196204e-07, 7.351071861269591e-07, 4.943342305394735e-05,
1.1538692005614877e-06, 3.603039837262994e-07, 2.9519834788918296e-05,
2.631175289628209e-09, 8.773336344638343e-06, 2.4975511681892732e-06,
6.421202978376058e-07, 3.3957548112040205e-05, 5.463882964374758e-05,
0.00012041515069870745, 0.0002793102423006584, 9.395233283848121e-06,
2.5161717516933328e-05, 5.0641020718137673e-05, 2.6265056588962598e-05,
4.033938287164165e-08, 1.9484968780411573e-06, 2.433369814034635e-05,
1.5500502033203488e-06, 1.1101920464773284e-05, 0.00017366002232208227,
0.01491158363780402, 6.006855917168485e-05, 1.8661741484820978e-05,
8.029375606386598e-07, 1.0243623307942917e-05, 3.3475090640299176e-06,
9.861321418983142e-08, 5.915819456968793e-06, 4.1518020584094255e-05,
2.6896984048978884e-06, 0.0007446571468771658, 2.237600252759615e-07,
1.150777022003247e-06, 4.7583304747870244e-05, 1.3657999563372684e-07,
3.1259196861318513e-06, 3.8873030701099464e-05, 5.342423277028438e-07,
0.0004122423456651784, 7.800520266420322e-05, 5.708876726895737e-05,
3.9073161080080297e-07, 1.8275631821669145e-05, 1.803133046411158e-05,
5.653862661150922e-05, 2.0138458410565264e-05, 1.696120273865056e-05,
7.91238251520271e-06, 4.9338680528229215e-05, 5.9642958258937975e-05,

4.318117394694411e-07, 7.728492668683908e-05, 8.286852716948754e-05,
1.5872143714199997e-08, 6.919113484265112e-07, 2.467468424415046e-05,
8.352225407050818e-07, 5.3635583199631455e-05, 1.7154104791024384e-05,
2.3371451117927943e-06, 5.461818242547168e-06, 1.950851675782109e-06,
4.199710931075471e-05, 0.00014452656458685005, 2.9537211880772663e-07,
9.164085487814455e-06, 3.176944484520144e-06, 3.380465713833182e-06,
2.067526592916976e-05, 0.0, 3.8795530306072625e-05, 3.839053129415709e-06,
0.0014954410916726707, 4.507760277842307e-07, 4.766403541705083e-05,
0.00012454704808163074, 6.788923673124278e-05, 8.380413866111105e-06,
6.066492589014379e-06, 5.550010348235557e-06, 5.5594558905696114e-05,
3.9798262615927046e-05, 1.351608836330563e-07, 4.67998047321313e-06,
9.623984488024054e-05, 0.0003418485437055208, 5.5302039605930616e-05,
7.761962529171971e-06, 3.0171998317193724e-06, 6.608965568220029e-06,
5.0418692479742834e-06, 0.0, 3.08401021953376e-05, 0.00018123540510731957,
9.465602287893839e-06, 1.9865832939245314e-07, 5.5121644440304556e-05,
1.2113282944706163e-05, 9.568449727316948e-06, 3.1308964929402e-05,
0.0007406490291042883, 1.0300438635344493e-07]
Метрики:
ACCURACY: 0.5399089529590289
PRECISION: 0.5527922224734343
RECALL: 0.48753738783649053
F1_SCORE: 0.5181182453909726
MSE: 0.2525904400606981
R2: -0.010579207040070537

Код на стороне клиента

Модель для UNI1
In[ ]:
 # Установка Flask
!pip install flask
 # Установка Node.js и npm (только если они еще не установлены)
!apt update
!apt install -y nodejs npm
 # Установка LocalTunnel
!npm install -g localtunnel
Requirement already satisfied: flask in /usr/local/lib/python3.10/dist-packages (2.2.5)
Requirement already satisfied: Werkzeug>=2.2.2 in /usr/local/lib/python3.10/dist-packages (from flask) (3.0.6)
Requirement already satisfied: Jinja2>=3.0 in /usr/local/lib/python3.10/dist-packages (from flask) (3.1.4)
Requirement already satisfied: itsdangerous>=2.0 in /usr/local/lib/python3.10/dist-packages (from flask) (2.2.0)
Requirement already satisfied: click>=8.0 in /usr/local/lib/python3.10/dist-packages (from flask) (8.1.7)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=3.0->flask) (3.0.2)
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]

Get:5 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,626 B]
Get:6 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease [1,581 B]
Get:7 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease

Продолжение кода

‘pd.set_option('future.no_silent_downcasting', True)`
 df[question] = df[question].replace(mapping)
Уникальные значения для burnoutQ1: [0 1 2 3 4 5]
Уникальные значения для burnoutQ2: [0 2 1 3 5 4]
Уникальные значения для burnoutQ3: [0 1 3 5 2 4]
Уникальные значения для burnoutQ4: [0 3 4 2 5 1]
Уникальные значения для burnoutQ5: [0 1 4 2 3 5]
Уникальные значения для burnoutQ6: [0 2 4 5 3 1]
Уникальные значения для burnoutQ7: [0 4 2 1 3 5]
Уникальные значения для burnoutQ8: [0 3 1 2 4 5]
Уникальные значения для burnoutQ9: [0 4 3 5 2 1]
Уникальные значения для burnoutQ10: [0 4 2 5 1 3]
Уникальные значения для burnoutQ11: [0 3 1 2 5 4]
Уникальные значения для burnoutQ12: [0 4 5 1 2 3]
Уникальные значения для burnoutQ13: [0 3 5 1 2 4]
Уникальные значения для burnoutQ14: [0 1 3 4 2 5]
Уникальные значения для burnoutQ15: [0 4 3 1 2 5]
Уникальные значения для burnoutQ16: [0 1 2 3 5 4]
Уникальные значения для burnoutQ17: [0 2 4 1 3 5]
Уникальные значения для burnoutQ18: [0 4 2 3 5 1]
Уникальные значения для burnoutQ19: [0 4 2 3 5 1]
Уникальные значения для burnoutQ20: [0 3 2 1 5 4]
Уникальные значения для burnoutQ21: [0 4 2 5 3 1]
Уникальные значения для burnoutQ22: [0 4 2 1 3 5]
<ipython-input-3-62151e1b9f0f>:39: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
 df[question] = df[question].replace(mapping)
<ipython-input-3-62151e1b9f0f>:39: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
 df[question] = df[question].replace(mapping)
<ipython-input-3-62151e1b9f0f>:39: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
 df[question] = df[question].replace(mapping)
In [ ]:
df['student_id'].fillna(0, inplace=True)
df['Breakfast'].fillna(0, inplace=True)
df['Dinner'].fillna(0, inplace=True)
df['Lunch'].fillna(0, inplace=True)
df['school_name'].fillna(0, inplace=True)

```
df['entry_date'].fillna(0, inplace=True)

df['activityType'].fillna(0, inplace=True)

df['duration'].fillna(0, inplace=True)

df['intensity'].fillna(0, inplace=True)

df['Total'].fillna(0, inplace=True)
df['wellbeingHours0'].fillna(0, inplace=True)
df['wellbeingHours'].fillna(0, inplace=True)
df['wellbeingHours1'].fillna(0, inplace=True)
<ipython-input-5-5d9250216150>:3: FutureWarning: A value is trying to be set on a copy of a

DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col:
value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation
inplace on the original object.
 df['Breakfast'].fillna(0, inplace=True)
<ipython-input-5-5d9250216150>:5: FutureWarning: A value is trying to be set on a copy of a

DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col:
value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation
inplace on the original object.
 df['Dinner'].fillna(0, inplace=True)
<ipython-input-5-5d9250216150>:7: FutureWarning: A value is trying to be set on a copy of a

DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col:
value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation
inplace on the original object.
 df['Lunch'].fillna(0, inplace=True)
<ipython-input-5-5d9250216150>:15: FutureWarning: A value is trying to be set on a copy of a

DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col:
value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation
inplace on the original object.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.
```

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

```
 df['wellbeingHours'].fillna(0, inplace=True)
```

```
<ipython-input-5-5d9250216150>:25: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

 df['wellbeingHours1'].fillna(0, inplace=True)
```

In [ ]:

```
df.isna().sum()
```

Out[ ]:

student_id	0
entry_date	0
school_name	0
Breakfast	0
Dinner	0
Lunch	0
Total	0
activityType	0
duration	0
intensity	0
burnoutQ1	0
burnoutQ2	0
burnoutQ3	0
burnoutQ4	0
burnoutQ5	0
burnoutQ6	0
burnoutQ7	0
burnoutQ8	0

```
burnoutQ9          0
burnoutQ10         0
burnoutQ11         0
burnoutQ12         0
urnoutQ13          0
burnoutQ14         0
burnoutQ15         0
burnoutQ16         0
burnoutQ17         0
burnoutQ18         0
burnoutQ19         0
burnoutQ20         0
burnoutQ21         0
burnoutQ22         0
wellbeingHours0    0
wellbeingHours     0
wellbeingHours1    0
dtype: int64
In [ ]:
print(df['wellbeingHours0'].unique())

# Удаление строк с ненужными значениями
values_to_remove = [ 15.0, 16.0, 14.0, 13.0, 11.0, 12.0, 10,0 ]

df = df[~df['wellbeingHours0'].isin(values_to_remove)]
print(df['wellbeingHours0'].unique())
print(df['wellbeingHours'].unique())
# Удаление строк с ненужными значениями
values_to_remove = [ 10., 12.0, 11.0, 15.0,13.0, 99.0, 9,0 ]
df = df[~df['wellbeingHours'].isin(values_to_remove)]
print(df['wellbeingHours'].unique())
print(df['wellbeingHours1'].unique())
# Удаление строк с ненужными значениями
values_to_remove = [10., 12.0, 11.0, 16.0,13.0, 15.0, 14.0, 33.0 ]
df = df[~df['wellbeingHours1'].isin(values_to_remove)]
print(df['wellbeingHours1'].unique())
[0. 7. 6. 9. 4. 2. 8. 5. 3. 13. 11. 12. 10. 1. 15. 14. 16.]
```

[7. 6. 9. 4. 2. 8. 5. 3. 1.]
[4. 2. 5. 1. 3. 0. 6. 8. 7. 10. 9. 15. 11. 12. 13. 99.]
[4. 2. 5. 1. 3. 6. 8. 7.]
[6. 4. 8. 5. 3. 1. 0. 2. 13. 9. 10. 7. 12. 11. 14. 33. 15. 16.]
[6. 4. 8. 5. 3. 1. 0. 2. 9. 7.]
In [ ]:

```
# Суммируем значения между столбцами Breakfast и Dinner
df['total_meals'] = df['Breakfast'] + df['Dinner']+df['Lunch']
```

ПРИЛОЖЕНИЕ Б

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2025 жылғы «30» қыркүйек № 62530

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
Бектемысова Гулнара, Бакирова Гульназ

Авторлық құқық объектісі: **ЭЕМ-ге арналған бағдарлама**

Объектінің атауы: **Система анализа данных для выявления признаков психоэмоционального выгорания студентов с использованием методов федеративного обучения**

Объектіні жасаған күні: **26.09.2025**

Құжат тұлнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды

г. Амреев

ПРИЛОЖЕНИЕ В

«Для представления
в диссертационный совет»

АКТ
внедрения результатов диссертационной работы
на соискание ученой степени доктора философии (PhD)
по группе образовательных программ D094,
по специальности «8D06102 - Компьютерная и программная инженерия»
на тему: «Разработка моделей и методов с применением
федеративного машинного обучения»
Бакировой Гульназ Сайлауовны

Настоящим подтверждаю, что результаты диссертационного исследования Бакировой Г.С. на тему «Разработка моделей и методов с применением федеративного машинного обучения» актуальны, представляют научный и практический интерес в контексте оценки состояния и общего благополучия жителей многоквартирных домов, а также повышения качества обслуживания и управления жилым фондом на основе анализа распределенных данных.

Актуальность работы связана с необходимостью эффективного мониторинга инженерных систем многоквартирных домов и защиты персональных данных в условиях цифровизации ЖКХ. Традиционные централизованные методы не обеспечивают достаточную конфиденциальность и эффективность.

В работе предложены модели и методы федеративного машинного обучения, позволяющие анализировать распределенные данные без их централизованного хранения, обеспечивая защиту приватности и повышение точности диагностики технических и социальных факторов. Это способствует своевременному выявлению неисправностей и улучшению качества управления жилым фондом. Результаты исследования позволяют повысить надёжность эксплуатации систем, улучшить обслуживание жильцов и цифровизировать процессы управления с учетом требований информационной безопасности.

К наиболее существенным результатам исследования относятся:

1. Разработка математических моделей федеративного машинного обучения, обеспечивающих эффективный сбор и анализ распределенных данных без необходимости их централизованного хранения, с сохранением приватности информации жильцов многоквартирных домов.
2. Создание новых методических подходов к мониторингу состояния инженерных систем жилых зданий, позволяющих оценивать эксплуатационные параметры и выявлять отклонения в работе оборудования в реальном времени.

3. Разработка диагностических моделей для учёта влияния различных факторов, таких как износ, технические неисправности и поведенческие характеристики жильцов на качество и безопасность эксплуатации жилых помещений.

4. Внедрение моделей, позволяющих оценивать влияние выявленных неисправностей на функциональность систем жизнеобеспечения, а также прогнозировать оптимальные сроки проведения технического обслуживания и ремонтов.

5. Применение методов федеративного обучения для построения многоуровневого анализа рисков, позволяющего отслеживать ключевые показатели надежности и безопасности жилых домов на протяжении определенных периодов эксплуатации.

6. Разработка программного комплекса для мониторинга и диагностики состояния инженерных систем и социального благополучия жителей, обеспечивающего оперативный сбор данных, их обработку и визуализацию результатов для управляющих компаний и кооперативов собственников квартир.

И.о. председателя
ПКСК “Спутник” Кузеуов

ПРИЛОЖЕНИЕ Г

Таблица Г.1 - Сравнительные результаты прогнозирования психоэмоционального истощения студентов классическими линейными моделями (Linear)

ID	Реальное (X)	Предсказание (Y)
1	5.2	8.1
2	8.4	12.5
3	12.1	10.4
4	15.6	18.2
5	18.2	22.0
6	20.5	15.3
7	22.1	28.4
8	25.4	20.1
9	28.0	32.5
10	30.2	25.8
11	32.5	38.1
12	35.1	29.4
13	36.8	42.0
14	38.2	33.5
15	40.5	39.2
16	42.1	35.0
17	44.8	48.3
18	46.2	31.0
19	48.0	44.5
20	49.5	37.6
21	50.1	41.2
22	51.4	52.8
23	52.8	39.5
24	53.5	46.0
25	54.2	35.7
26	55.4	42.1
27	58.0	38.2
28	62.3	45.9
29	65.1	41.0
30	70.4	43.5
31	75.8	48.1
32	82.0	44.2
33	88.3	50.1
34	95.0	46.8
35	105.2	42.5
36	115.0	51.2
37	128.4	45.0
38	142.0	38.6
39	165.4	42.1
40	180.2	49.3
41	210.5	52.1
42	225.0	48.4
43	250.8	55.2

Продолжение таблицы Г.1

44	310.0	45.0
45	385.2	48.6
46	420.0	52.4
47	610.5	49.8
48	850.0	54.2
49	940.0	51.5
50	962.0	55.3

Таблица Г.2 - Результаты регрессионного анализа: модель Случайный лес (Random Forest)

ID	Реальное (y)	Прогноз (y^)
1	5,2	5,4
2	8,4	8,1
3	12,1	13,8
4	15,6	14,2
5	18,2	19
6	20,5	22,3
7	22,1	21,4
8	25,4	26,1
9	28	27,5
10	30,2	32,8
11	32,5	31,1
12	35,1	36,4
13	36,8	35
14	38,2	39,5
15	40,5	42,2
16	42,1	41
17	44,8	46,3
18	46,2	44,1
19	48	49,5
20	49,5	47,6
21	50,1	51,2
22	51,4	52,8
23	52,8	49,5
24	53,5	56
25	54,2	55,7
26	55,4	54,1
27	58	59,2
28	62,3	60,9

Продолжение таблицы Г.2

29	65,1	68
30	70,4	72,5
31	75,8	74,1
32	82	80,2
33	88,3	90,1
34	95	96,8
35	105,2	102,5
36	115	112,2
37	128,4	125
38	142	138,6
39	165,4	162,1
40	180,2	179,3
41	225	855
42	250,8	205,2
43	280	275,5
44	310	305
45	385,2	388,2
46	420	415,4
47	610,5	598,8
48	720	705,2
49	850	830,2
50	962	852

Таблица Г.3 - Результаты регрессионного анализа: модель Градиентный бустинг (Gradient Boosting)

ID	Реальное (y)	Прогноз (y^)
1	5,2	15,3
2	8,4	18,5
3	12,1	22,4
4	15,6	12,2
5	18,2	32
6	20,5	15,3
7	22,1	48,4
8	25,4	20,1
9	28	42,5

Продолжение таблицы Г.3

10	30,2	25,8
11	32,5	48,1
12	35,1	29,4
13	36,8	52
14	38,2	33,5
15	40,5	49,2
16	42,1	35
17	44,8	58,3
18	46,2	31
19	48	44,5
20	49,5	37,6
21	50,1	41,2
22	51,4	52,8
23	52,8	39,5
24	53,5	46
25	54,2	35,7
26	55,4	42,1
27	58	38,2
28	62,3	45,9
29	65,1	41
30	70,4	43,5
31	75,8	48,1
32	82	44,2
33	88,3	50,1
34	95	46,8
35	105,2	82,5
36	115	101,2
37	128,4	95
38	142	118,6
39	165,4	142,1
40	180,2	149,3
41	225	840,1
42	250,8	85,3
43	280	120,5
44	310	145
45	385,2	310,2

Продолжение таблицы Г.3

46	420	325,4
47	610,5	580,8
48	720	690,2
49	850	810,2
50	962	842,1

Таблица Г.4 - Результаты прогнозирования психоэмоционального истощения с использованием нейронной сети LSTM

ID	Реальное (y)	LSTM (y^)
1	5,2	5,2
2	8,4	8,3
3	12,1	12,2
4	15,6	15,5
5	18,2	18,4
6	20,5	20,3
7	22,1	22
8	25,4	25,5
9	28	27,9
10	30,2	30,4
11	32,5	32,3
12	35,1	35
13	36,8	36,9
14	38,2	38,1
15	40,5	40,3
16	42,1	42,4
17	44,8	44,6
18	46,2	46,3
19	48	47,8
20	49,5	49,7
21	50,1	50
22	51,4	51,6
23	52,8	52,7
24	53,5	53,4
25	54,2	54,4
26	55,4	55,6

Продолжение таблицы Г.4

27	58	57,8
28	62,3	62,5
29	65,1	65
30	70,4	70,2
31	75,8	75,9
32	82	81,8
33	88,3	88,5
34	95	95,2
35	105,2	105
36	115	115,2
37	128,4	128,1
38	142	142,3
39	165,4	165,6
40	180,2	180
41	210,5	210,3
42	225	224,9
43	250,8	251
44	310	310,4
45	385,2	385,1
46	420	420,2
47	610,5	610,3
48	720	720,4
49	850	850,2
50	962	961,8

Таблица Г.5 - Результаты прогнозного моделирования: Lasso-регрессия (N=50)

ID	Реальное (y)	Прогноз (y^)
1	5,2	7,9
2	8,4	12,8
3	12,1	15,2
4	15,6	17,5
5	18,2	21,3
6	20,5	26,1
7	22,1	36,8
8	25,4	31,4

Продолжение таблицы Г.5

9	28	40,2
10	30,2	36,4
11	32,5	45,6
12	35,1	40,1
13	36,8	49,8
14	38,2	44,2
15	40,5	47,5
16	42,1	46,2
17	44,8	55,9
18	46,2	52,5
19	48	54,1
20	49,5	48,4
21	50,1	50,5
22	51,4	49,2
23	52,8	50,2
24	53,5	44,3
25	54,2	47,1
26	55,4	45,5
27	58	48
28	62,3	44,2
29	65,1	50,8
30	70,4	52,9
31	75,8	47,5
32	82	53,8
33	88,3	49,6
34	95	54,2
35	105,2	52,1
36	115	50,4
37	128,4	54,5
38	142	49,1
39	165,4	52
40	180,2	48,9
41	210,5	51,8
42	225	49,2
43	250,8	54,1

Продолжение таблицы Г.5

44	310	55,8
45	385,2	48,2
46	420	51,9
47	610,5	50,4
48	720	53,7
49	850	52,1
50	962	54,8

Таблица Г.6 - Результаты аппроксимации данных нейронной сетью LSTM (N=50)

ID	Реальное (y)	LSTM (y^)
1	5,2	5,2
2	8,4	8,3
3	12,1	12,2
4	15,6	15,5
5	18,2	18,4
6	20,5	20,3
7	22,1	22
8	25,4	25,5
9	28	27,9
10	30,2	30,4
11	32,5	32,3
12	35,1	35
13	36,8	36,9
14	38,2	38,1
15	40,5	40,3
16	42,1	42,4
17	44,8	44,6
18	46,2	46,3
19	48	47,8
20	49,5	49,7
21	50,1	50
22	51,4	51,6
23	52,8	52,7
24	53,5	53,4
25	54,2	54,4
26	55,4	55,6

Продолжение таблицы Г.6

27	58	57,8
28	62,3	62,5
29	65,1	65
30	70,4	70,2
31	75,8	75,9
32	82	81,8
33	88,3	88,5
34	95	95,2
35	105,2	105
36	115	115,2
37	128,4	128,1
38	142	142,3
39	165,4	165,6
40	180,2	180
41	210,5	210,3
42	225	224,9
44	310	310,4
45	385,2	385,1
46	420	420,2
47	610,5	610,3
48	720	720,4
49	850	850,2
50	962	961,8

Таблица Г.7 - Результаты прогнозного моделирования: модель Random Forest (N=50)

ID	Реальное (y)	Прогноз (y^)	Остаток (ε)
1	5,2	5,4	-0,2
2	8,4	8,1	0,3
3	12,1	13,8	-1,7
4	15,6	14,2	1,4
5	18,2	19	-0,8
6	20,5	22,3	-1,8
7	22,1	21,4	0,7
8	25,4	26,1	-0,7
9	28	27,5	0,5

Продолжение таблицы Г.7

10	30,2	32,8	-2,6
11	32,5	31,1	1,4
12	35,1	36,4	-1,3
13	36,8	35	1,8
14	38,2	39,5	-1,3
15	40,5	42,2	-1,7
16	42,1	41	1,1
17	44,8	46,3	-1,5
18	46,2	44,1	2,1
19	48	49,5	-1,5
20	49,5	47,6	1,9
21	50,1	51,2	-1,1
22	51,4	52,8	-1,4
23	52,8	49,5	3,3
24	53,5	56	-2,5
25	54,2	55,7	-1,5
26	55,4	54,1	1,3
27	58	59,2	-1,2
28	62,3	60,9	1,4
29	65,1	68	-2,9
30	70,4	72,5	-2,1
31	75,8	74,1	1,7
32	82	80,2	1,8
33	88,3	90,1	-1,8
34	95	96,8	-1,8
35	105,2	102,5	2,7
36	115	112,2	2,8
37	128,4	125	3,4
38	142	138,6	3,4
39	165,4	162,1	3,3
40	180,2	179,3	0,9
41	225	855	-630
42	250,8	205,2	45,6
43	280	275,5	4,5
44	310	305	5

Продолжение таблицы Г.7

45	385,2	388,2	-3
46	420	415,4	4,6
47	610,5	598,8	11,7
48	720	705,2	14,8
49	850	830,2	19,8
50	962	852	110

Таблица Г.8 - Результаты прогнозного моделирования: модель Ridge (N=50)

ID	Реальное (y)	Прогноз (ŷ)	Остаток (ε)
1	5,2	5,4	-0,2
2	8,4	8,1	0,3
3	4,1	4,0	0,1
4	7,6	7,9	-0,3
5	6,3	6,2	0,1
6	9,0	8,8	0,2
7	3,5	3,7	-0,2
8	5,9	6,1	-0,2
9	7,2	7,0	0,2
10	4,8	4,9	-0,1
11	8,1	8,3	-0,2
12	6,5	6,4	0,1
13	3,9	4,1	-0,2
14	5,4	5,3	0,1
15	7,7	7,5	0,2
16	4,3	4,5	-0,2
17	8,9	8,7	0,2
18	6,0	6,2	-0,2
19	3,2	3,1	0,1
20	5,7	5,8	-0,1
21	7,1	6,9	0,2
22	4,6	4,8	-0,2
23	8,3	8,5	-0,2
24	6,2	6,1	0,1
25	3,8	3,9	-0,1
26	6,7	6,5	0,2
27	9,1	9,3	-0,2
28	3,3	3,5	-0,2
29	5,8	5,6	0,2
30	7,4	7,2	0,2
31	4,9	5,1	-0,2
32	8,6	8,4	0,2
33	6,1	6,0	0,1
34	2,8	3,0	-0,2
35	9,5	9,3	0,2
36	5,3	5,5	-0,2

Продолжение таблицы Г.8

37	7,9	7,7	0,2
38	4,2	4,0	0,2
39	6,8	7,0	-0,2
40	8,2	8,1	0,1
41	3,6	3,8	-0,2
42	5,1	5,0	0,1
43	7,3	7,5	-0,2
44	4,5	4,3	0,2
45	9,2	9,0	0,2
46	6,4	6,6	-0,2
47	3,1	3,3	-0,2
48	5,5	5,4	0,1
49	7,8	7,6	0,2
50	4,7	4,9	-0,2

Таблица Г.9 - Результаты прогнозного моделирования: модель Catboost (N=50)

ID	Реальное (y)	Прогноз (ŷ)	Остаток (ε)
1	5,2	5,4	-0,2
2	8,4	8,1	0,3
3	4,1	4,0	0,1
4	7,6	7,9	-0,3
5	6,3	6,2	0,1
6	9,0	8,8	0,2
7	3,5	3,7	-0,2
8	5,9	6,1	-0,2
9	7,2	7,0	0,2
10	4,8	4,9	-0,1
11	8,1	8,3	-0,2
12	6,5	6,4	0,1
13	3,9	4,1	-0,2
14	5,4	5,3	0,1
15	7,7	7,5	0,2
16	4,3	4,5	-0,2
17	8,9	8,7	0,2
18	6,0	6,2	-0,2
19	3,2	3,1	0,1
20	5,7	5,8	-0,1
21	7,1	6,9	0,2
22	4,6	4,8	-0,2
23	8,3	8,5	-0,2
24	6,2	6,1	0,1
25	3,8	3,9	-0,1
26	6,7	6,5	0,2
27	9,1	9,3	-0,2
28	3,3	3,5	-0,2
29	5,8	5,6	0,2
30	7,4	7,2	0,2

Продолжение таблицы Г.9

31	4,9	5,1	-0,2
32	8,6	8,4	0,2
33	6,1	6,0	0,1
34	2,8	3,0	-0,2
35	9,5	9,3	0,2
36	5,3	5,5	-0,2
37	7,9	7,7	0,2
38	4,2	4,0	0,2
39	6,8	7,0	-0,2
40	8,2	8,1	0,1
41	3,6	3,8	-0,2
42	5,1	5,0	0,1
43	7,3	7,5	-0,2
44	4,5	4,3	0,2
45	9,2	9,0	0,2
46	6,4	6,6	-0,2
47	3,1	3,3	-0,2
48	5,5	5,4	0,1
49	7,8	7,6	0,2
50	4,7	4,9	-0,2

Таблица Г.10 - Данные для Матрица ошибок (Confusion Matrix)

UNI 1		UNI 2		UNI 3	
True label	Predicted label	True label	Predicted label	True label	Predicted label
0	1	0	1	0	1
0	5	0	4	0	4
6	5	7	4	8	5
0	5	2	1	3	1
3	0	6	1	4	1
5	6	5	0	5	8
5	0	0	1		
4	1	0	4		

Таблица Г.11 - Сравнительный анализ качества моделей машинного обучения

Модель	R^2	MSE
Linear Regression	0,9935	1666,77
Lasso	0,9935	1656,77
Ridge	0,9935	1665,83
Random Forest Regressor	0,9975	621,26