



Хакатон

Прогнозирование погоды с помощью нейросетвых моделей

Выполнили студенты группы М24-525

Груданов Н.А.

Лемцё В.В.



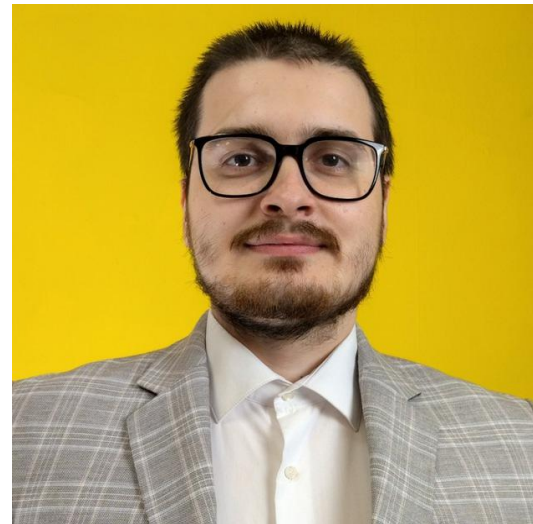
Команда

Название команды: Машинное заблуждение

Состав команды



Лемцё В.В.
ML development
Python developer



Груданов Н.А.
Лид команды
ML development
Data Analytics

Цели и задачи

Цель проекта:

Разработка интеллектуальной системы прогнозирования погоды с использованием современных технологий искусственного интеллекта и машинного обучения.

Задачи:

1. Анализ современных моделей для прогноза погоды
2. Поиск данных о погоде
3. Обучение моделей для прогноза
4. Интеграция моделей через LangChain
5. Подключение Gigachat для обработки естественного языка
6. Создание Telegram-бота для взаимодействия с пользователями

Данные

Источник данных - RP5.ru

- Разработан компанией "Расписание Погоды" (Санкт-Петербург) с 2004 года
- Имеет лицензию на деятельность в области гидрометеорологии
- Предоставляет данные для 173,000 населенных пунктов
- Использует данные с 9,800 метеостанций SYNOP и 5,500 станций METAR

Использованные данные:

- Метеостанция: Москва ВДНХ
- Период: 2005-2024 гг.

Описание

datetime - Дата и время
T - Температура воздуха
Po - Давление на уровне моря
P - Давление на станции
Pa - Изменение давления за 3 часа
U - Относительная влажность
DD - Направление ветра
Ff - Скорость ветра
ff10 - Скорость ветра (10 минут)
ff3 - Скорость ветра (3 минуты)
N - Облачность
WW - Погодные явления
W1 - Погодное явление 1
W2 - Погодное явление 2
Tn - Минимальная температура
Tx - Максимальная температура
Cl - Тип облаков
Nh - Количество облаков

H - Высота нижней границы облаков
Cm - Средний уровень облаков
Ch - Высокий уровень облаков
VV - Горизонтальная видимость
Td - Точка росы
RRR - Количество осадков
tR - Длительность осадков
E - Испарение
Tg - Температура почвы
E' - Испарение (альтернативное)
sss - Состояние почвы

Описание

#	Column	Non-Null Count	Dtype
0	datetime	67383 non-null	float64
1	T	67214 non-null	float64
2	Po	67253 non-null	float64
3	P	43830 non-null	float64
4	Pa	67346 non-null	float64
5	U	67340 non-null	object
6	DD	67340 non-null	float64
7	Ff	199 non-null	float64
8	ff10	1479 non-null	float64
9	ff3	63658 non-null	object
10	N	67390 non-null	object
11	WW	30831 non-null	object
12	W1	30831 non-null	object
13	W2	11713 non-null	float64
14	Tn	7233 non-null	float64

#	Column	Non-Null Count	Dtype
15	Tx	50844 non-null	object
16	Cl	53880 non-null	object
17	Nh	53904 non-null	object
18	H	32822 non-null	object
19	Cm	29374 non-null	object
20	Ch	41480 non-null	object
21	VV	67354 non-null	float64
22	Td	21555 non-null	object
23	RRR	21556 non-null	float64
24	tR	3099 non-null	object
25	E	2482 non-null	float64
26	Tg	2719 non-null	object
27	E'	2495 non-null	object
28	sss	0 non-null	float64



Пример данных

datetime	T	Po	P	Pa	U	DD	Ff	ff10	ff3
-1.3	721.8 100%.	735.4	-1.4	95.0	Ветер, дующий с юго-запада				1.0
-2.0	723.2 100%.	736.9	-3.6	93.0	Ветер, дующий с юго-юго-востока			3.0	
-1.3	726.8 100%.	740.6	-3.2	74.0	Ветер, дующий с юга			2.0	
-1.0	730.0 100%.	743.8	-1.8	80.0	Штиль, безветрие			0.0	
-0.9	731.8 100%.	745.6	-0.6	90.0	Ветер, дующий с западо-юго-запада				1.0

Пример данных

N WW W1

Снег непрерывный умеренный в срок наблюдения.

Снег и/или другие виды твердых осадков

Метель

Снег непрерывный сильный в срок наблюдения. Снег и/или другие виды твердых осадков Облака покрывали более половины неба в течение всего соответствующего периода.

Состояние неба в общем не изменилось. Снег и/или другие виды твердых осадков Облака покрывали более половины неба в течение всего соответствующего периода.

Состояние неба в общем не изменилось. Снег и/или другие виды твердых осадков Облака покрывали более половины неба в течение всего соответствующего периода.



Пример данных

Cl	Nh	H
100%.	300-600	Высокослоистые непросвечивающие или слоисто-дождевые.
100%.	300-600	Высокослоистые непросвечивающие или слоисто-дождевые.
100%.	600-1000	
100%.	300-600	



Анализ



Анализ

Проверка на null значений

datetime 7

T 176

Po 137

P 23560

Pa 44

U 50

DD 50

Ff 67191

ff10 65911

ff3 3732

N 0

WW 36559

W1 36559

W2 55677

Tn 60157

Tx 16546

Cl 13510

Nh 13486

H 34568

Cm 38016

Ch 25910

VV 36

Td 45835

RRR 45834

tR 64291

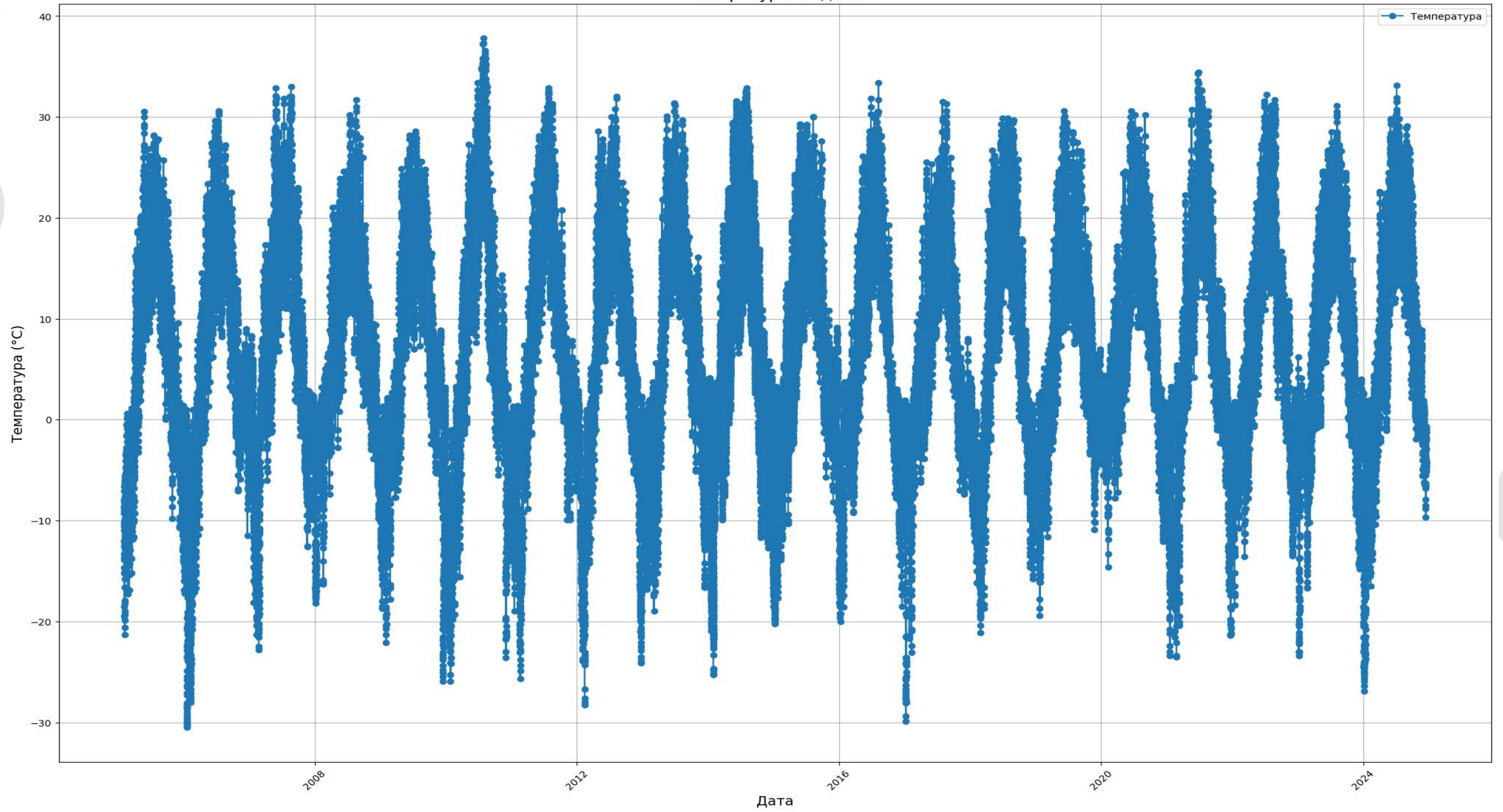
E 64908

Tg 64671

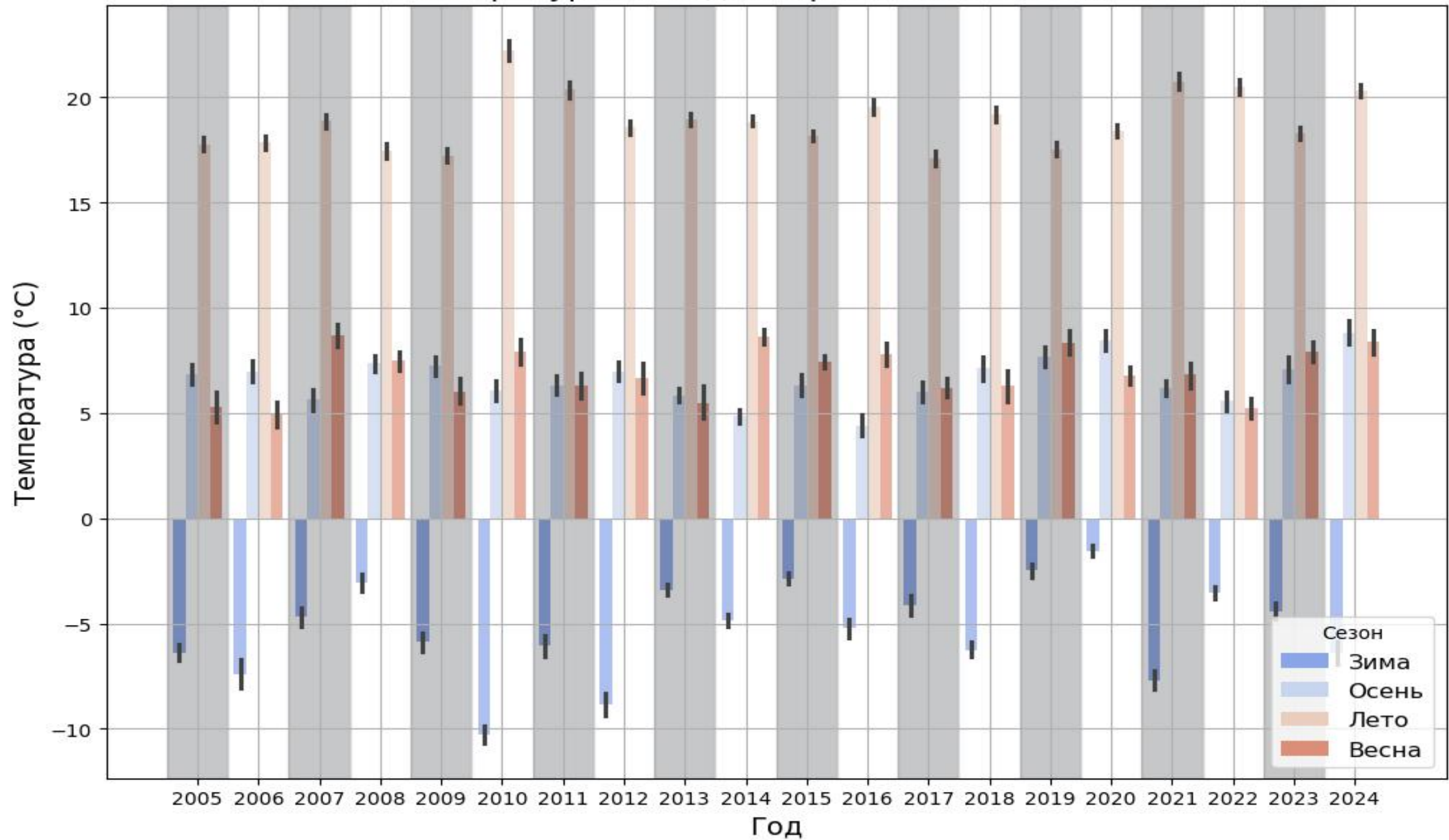
E' 64895

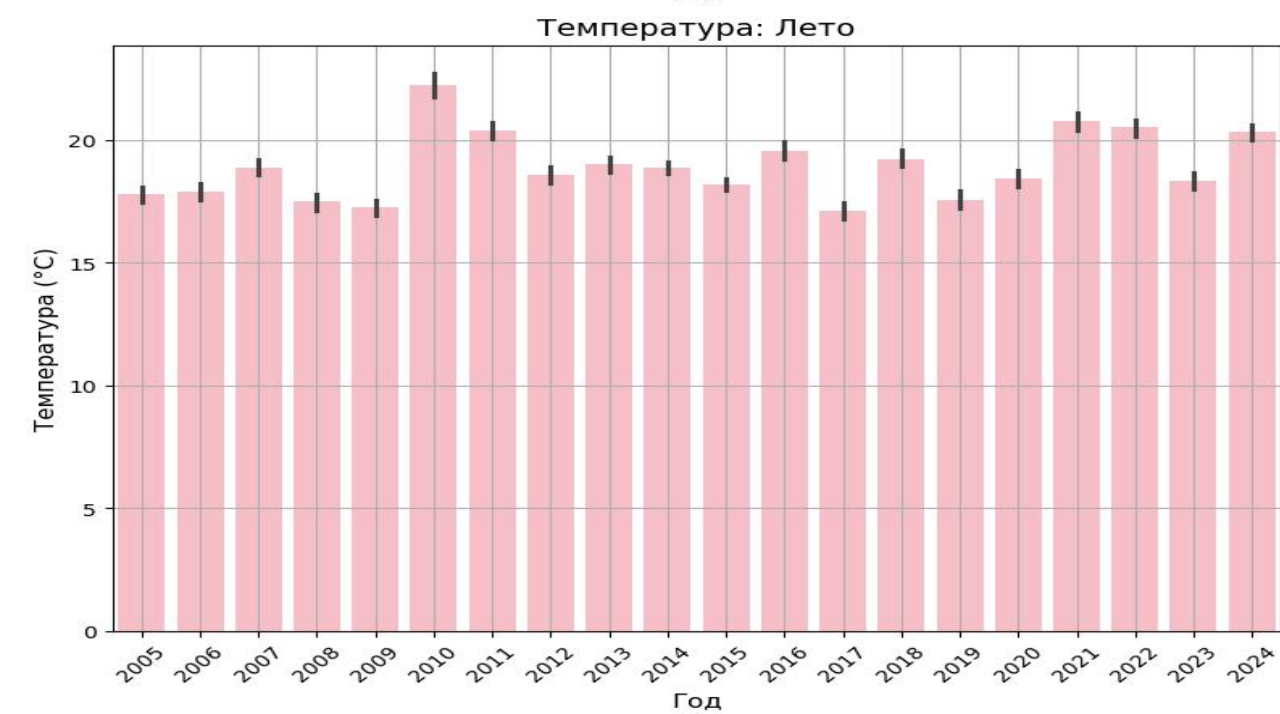
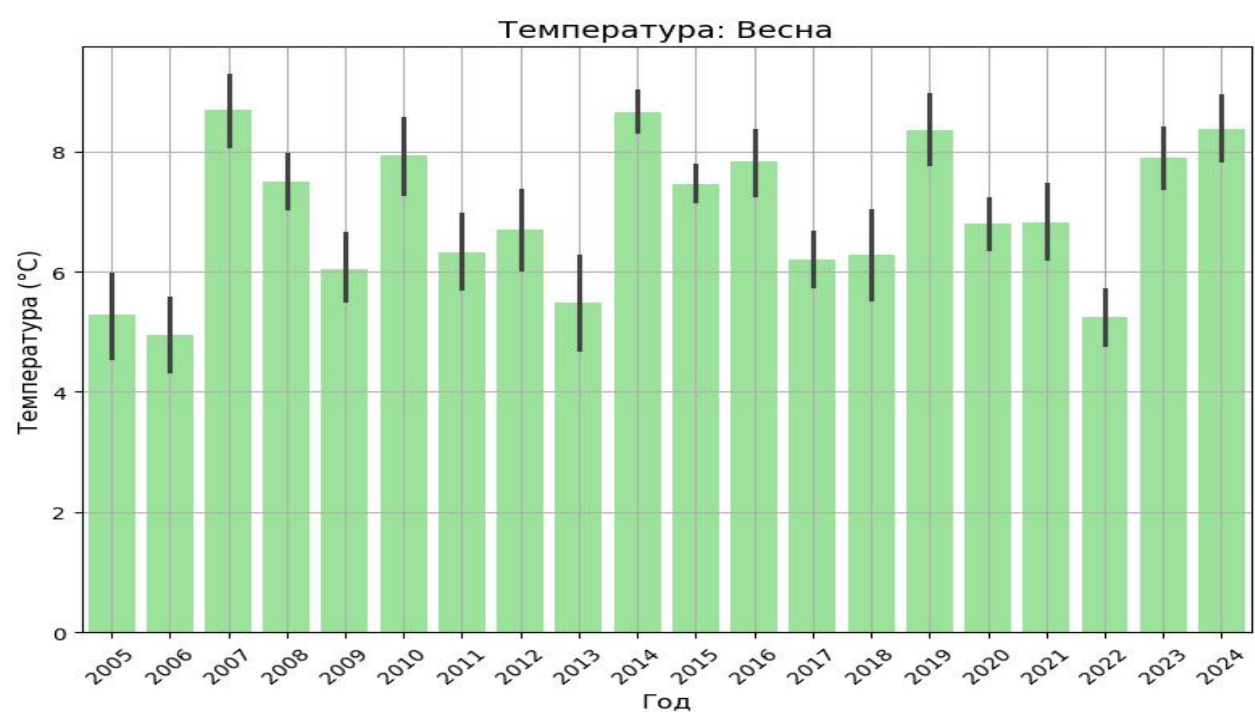
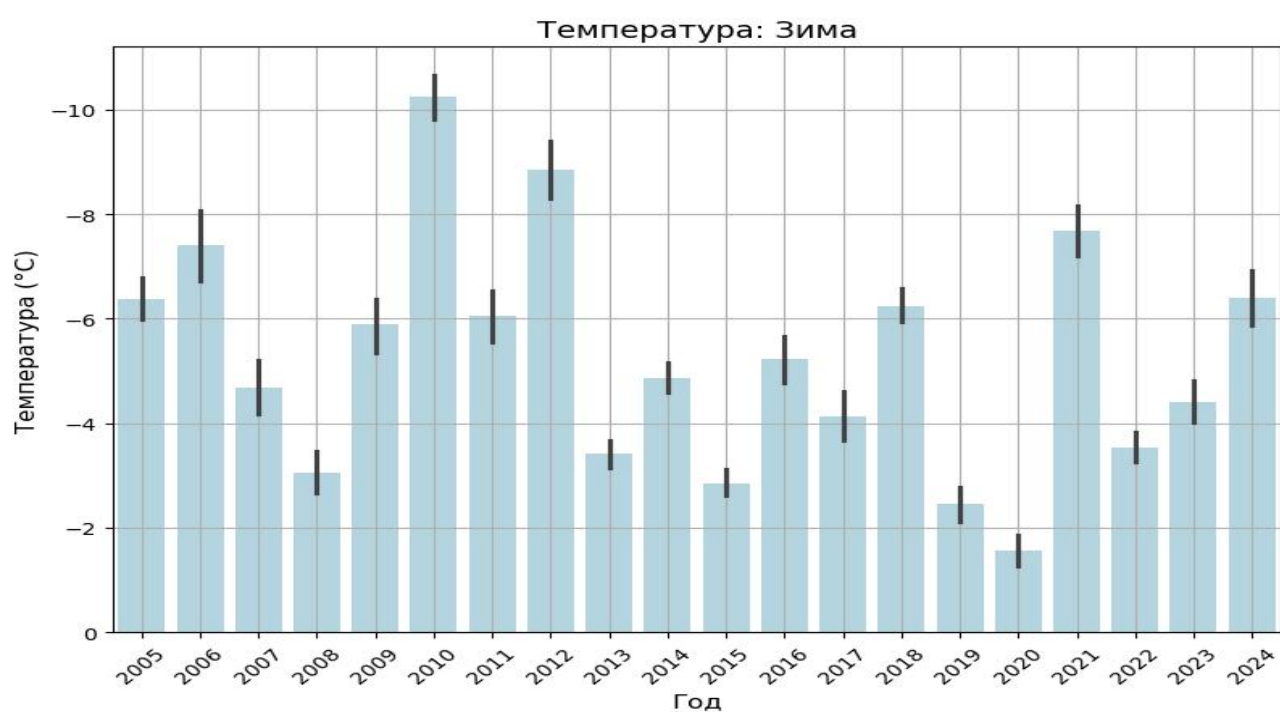
sss 67390

Температура по датам



Температура по годам с разбивкой на сезоны







Анализ

Проверим влияет ли день недели на температуру



Корреляция Спирмена

Корреляция Спирмена:

Коэффициент корреляции: -0.0012

P-значение: 0.7575

Близок к нулю, что указывает на отсутствие значимой монотонной связи между днем недели и температурой.

P-значение говорит о том, что связь между переменными статистически незначима.

ANOVA

ANOVA (анализ дисперсии) используется для проверки гипотезы о равенстве средних значений группы (в данном случае — температуры для разных дней недели).

F-статистика: 0.3547

P-значение: 0.9076

F-статистика (0.3547): Это отношение межгрупповой дисперсии к внутригрупповой дисперсии. Низкое значение F говорит о том, что различия между группами малы по сравнению с вариацией внутри групп.

P-значение (0.9076) следовательно, нет статистически значимых различий в температурах для разных дней недели.



Анализ

Проверим влияет ли время дня и ночи
на температуру

Анализ

Сезон: Зима

Количество

наблюдений: 16885

Корреляция: 0.0368

P-значение: 0.0000

Сезон: Весна

Количество

наблюдений: 17615

Корреляция: 0.1472

P-значение: 0.0000

Сезон: Лето

Количество наблюдений:

16743

Корреляция: 0.3680

P-значение: 0.0000

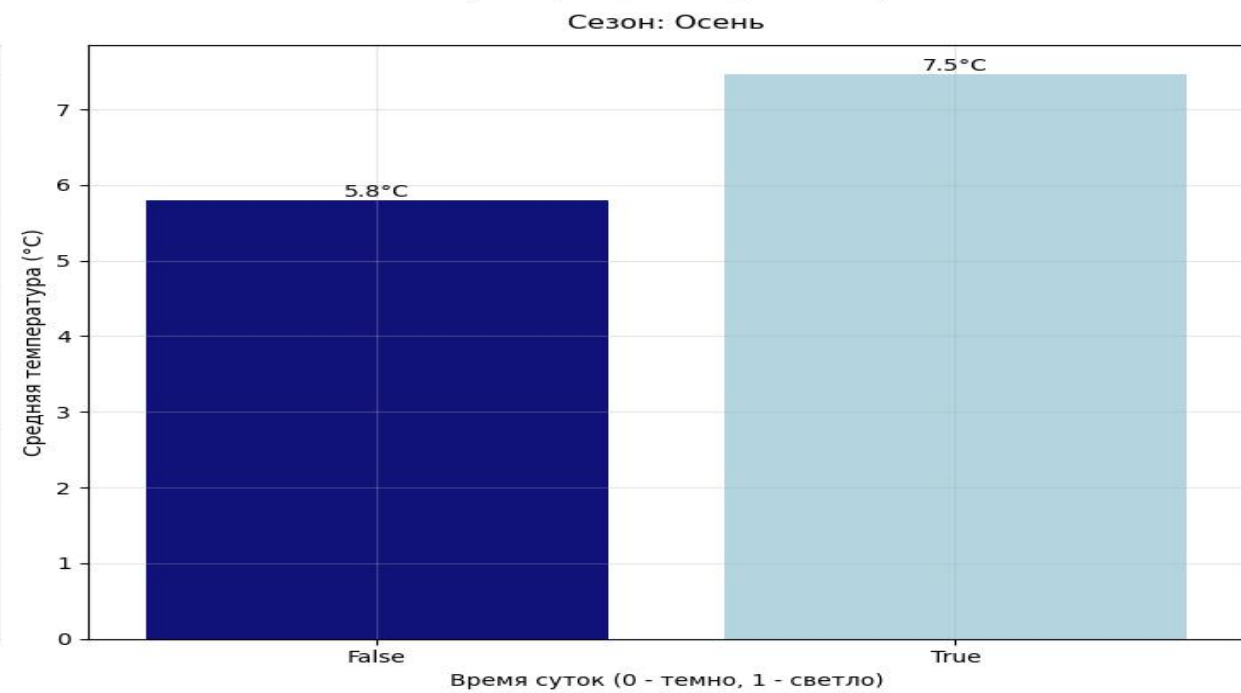
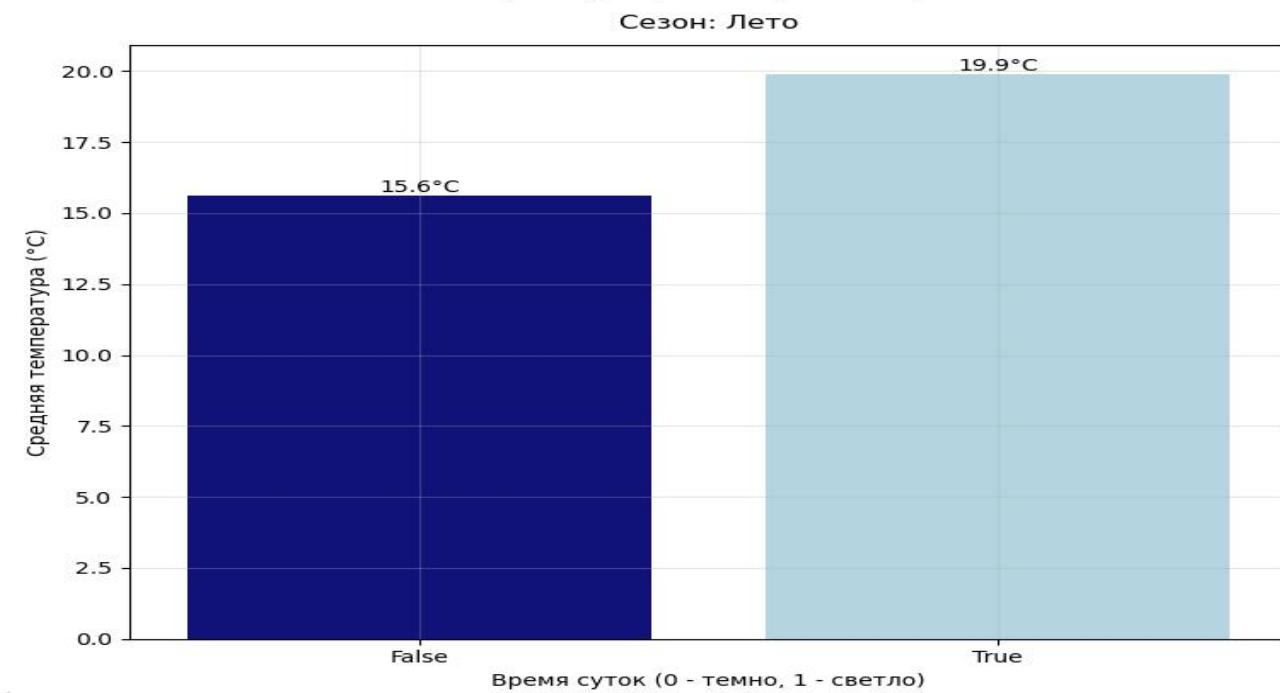
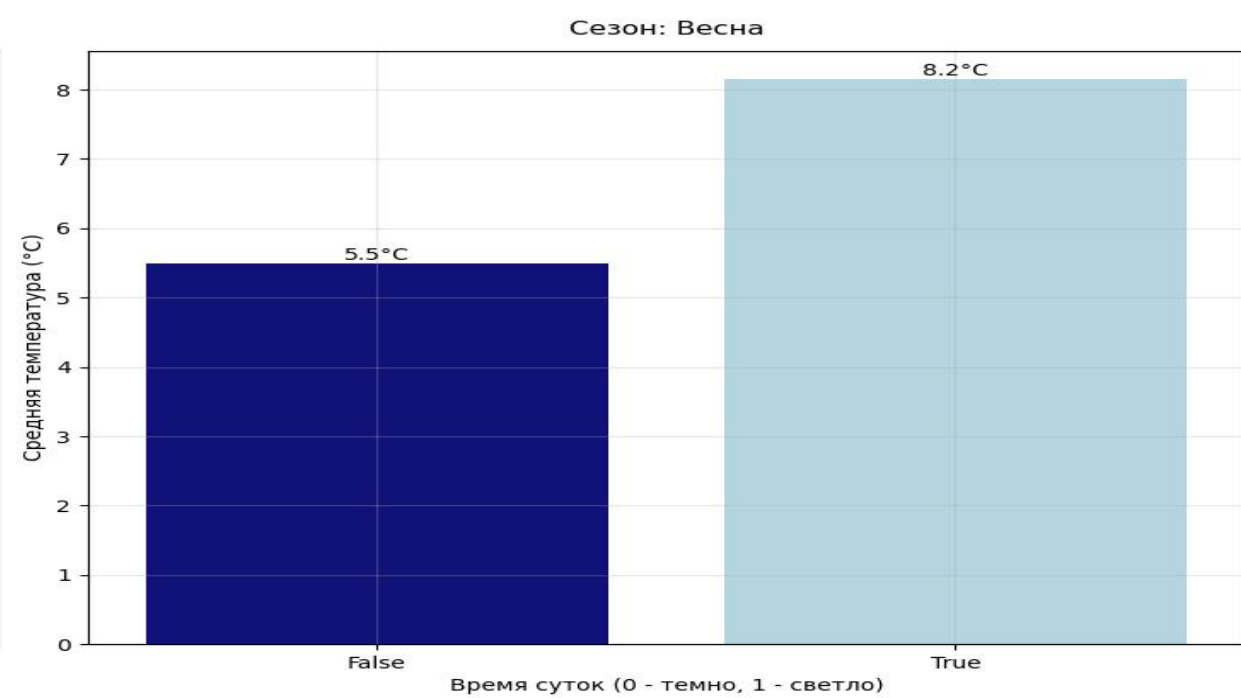
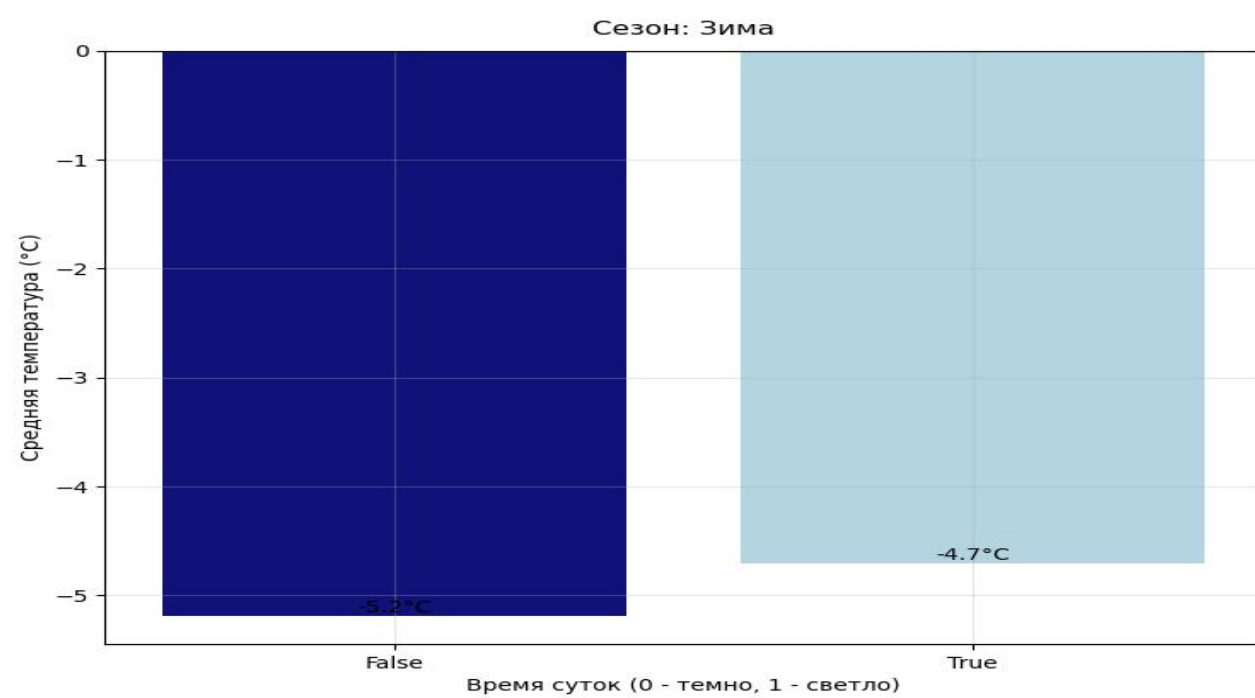
Сезон: Осень

Количество наблюдений:

16140

Корреляция: 0.1022

P-значение: 0.0000





Компоненты решения





Модель	Тип архитектуры	Преимущества	Недостатки	Применение
BERT	Трансформер	<ul style="list-style-type: none">- Хорошо работает с последовательностями- Учитывает контекст- Может обрабатывать пропуски в данных	<ul style="list-style-type: none">- Требуется много вычислительных ресурсов- Сложная архитектура	Краткосрочные и среднесрочные прогнозы
LSTM	Рекуррентная нейронная сеть	<ul style="list-style-type: none">- Эффективна для временных рядов- Хорошо запоминает долгосрочные зависимости- Относительно простая архитектура	<ul style="list-style-type: none">- Может быть медленной на длинных последовательностях	Краткосрочные прогнозы с высокой точностью
Random Forest	Ансамблевый метод	<ul style="list-style-type: none">- Устойчив к шумам- Не требует нормализации данных- Быстрое обучение	<ul style="list-style-type: none">- Ограниченная способность к экстраполяции- Может переобучаться	Базовые прогнозы и бенчмаркинг
GraphCast	Графовая нейронная сеть	<ul style="list-style-type: none">- Учитывает пространственные связи- Высокая точность для глобальных прогнозов	<ul style="list-style-type: none">- Очень требователен к ресурсам- Сложная инфраструктура	Глобальные среднесрочные прогнозы
Autoformer	Декомпозиционная архитектура	<ul style="list-style-type: none">- Эффективен для длительных прогнозов- На 38% точнее базовых моделей	<ul style="list-style-type: none">- Сложная настройка- Требуется качественных временных признаков	Долгосрочные прогнозы
PatchTST	Трансформер с патч-обработкой	<ul style="list-style-type: none">- Эффективная обработка длинных последовательностей- Меньше требований к памяти	<ul style="list-style-type: none">- Относительно новая модель- Меньше доступных реализаций	Долгосрочные прогнозы с оптимизацией памяти



Модели прогнозирования

- BERT модель (bert_weather_model)
 - Трансформер для анализа временных рядов
- LSTM модель (weather_lstm_model)
 - Долгая краткосрочная память для прогнозирования
- Random Forest (random_forest_model)
 - Ансамблевый метод для робастных прогнозов



Архитектура LSTM модели

Структура модели:

- Input size: 18 признаков (температура, сезонные признаки, освещенность)
- Hidden size: 64 нейрона
- Количество слоев: 2
- Batch size: 32
- Количество эпох: 100
- Оптимизатор: Adam с learning rate 0.001
- Функция потерь: MSE (Mean Squared Error)

Метрики качества модели

Основные показатели:

- MSE (Mean Squared Error): 1.5811
 - Показывает среднеквадратичную ошибку предсказания
 - Чем ближе к 0, тем лучше
- RMSE (Root Mean Square Error): 1.2574
 - Среднеквадратичное отклонение в градусах
 - Показывает среднюю ошибку прогноза в $\pm 1.26^{\circ}\text{C}$
- MAE (Mean Absolute Error): 0.8901
 - Средняя абсолютная ошибка
 - В среднем прогноз отклоняется на 0.89°C
- R^2 (Coefficient of Determination): 0.9864
 - Коэффициент детерминации
 - Модель объясняет 98.64% вариации в данных
 - Очень высокий показатель качества модели



Архитектура системы

Основные компоненты системы:

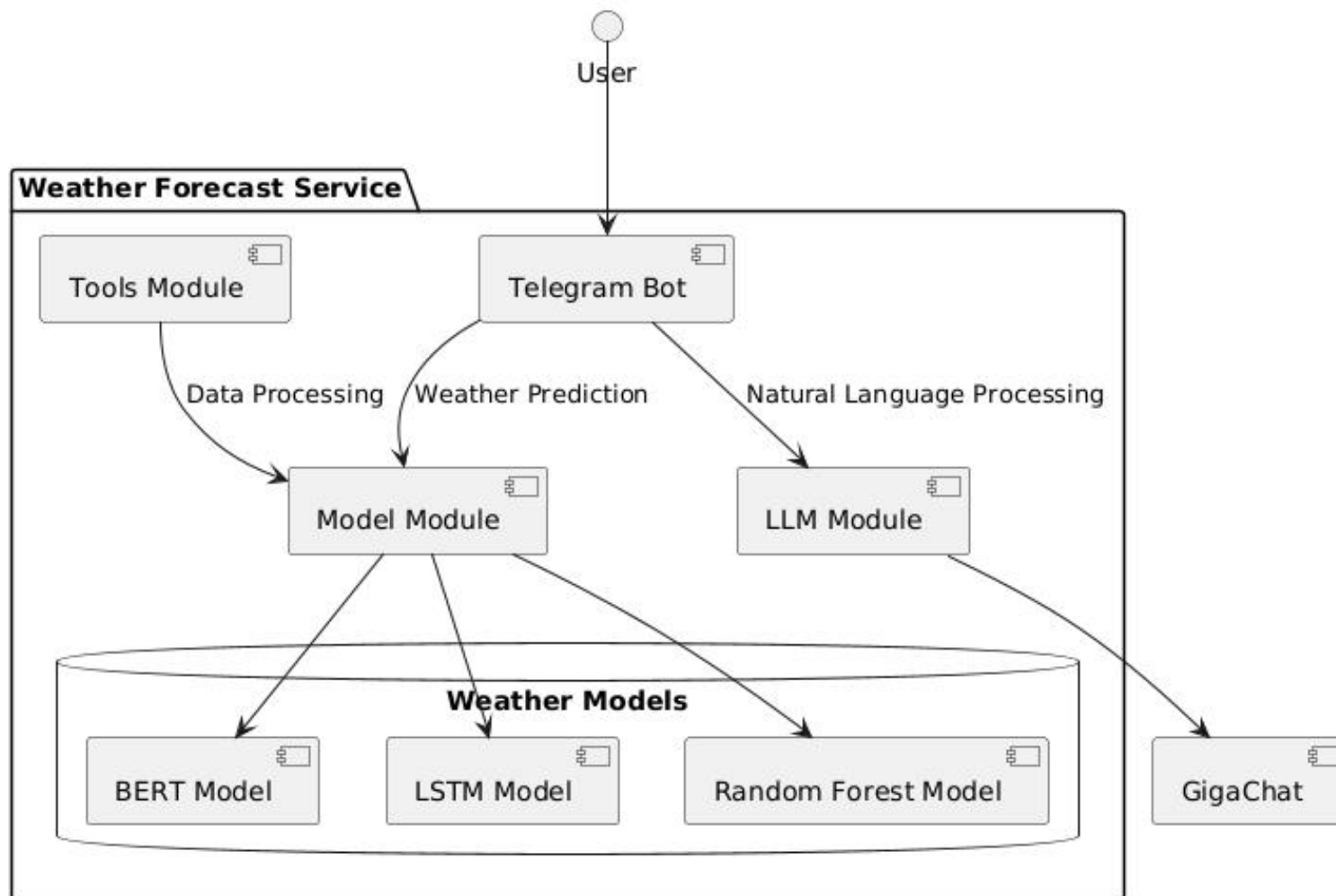
- Модуль прогнозирования (model)
 - Три типа моделей машинного обучения
 - Система выбора модели для прогноза
- Модуль обработки естественного языка (llm)
 - Интеграция с GigaChat-Max
 - Обработка пользовательских запросов
- Telegram бот (bot)
 - Пользовательский интерфейс
 - Маршрутизация запросов



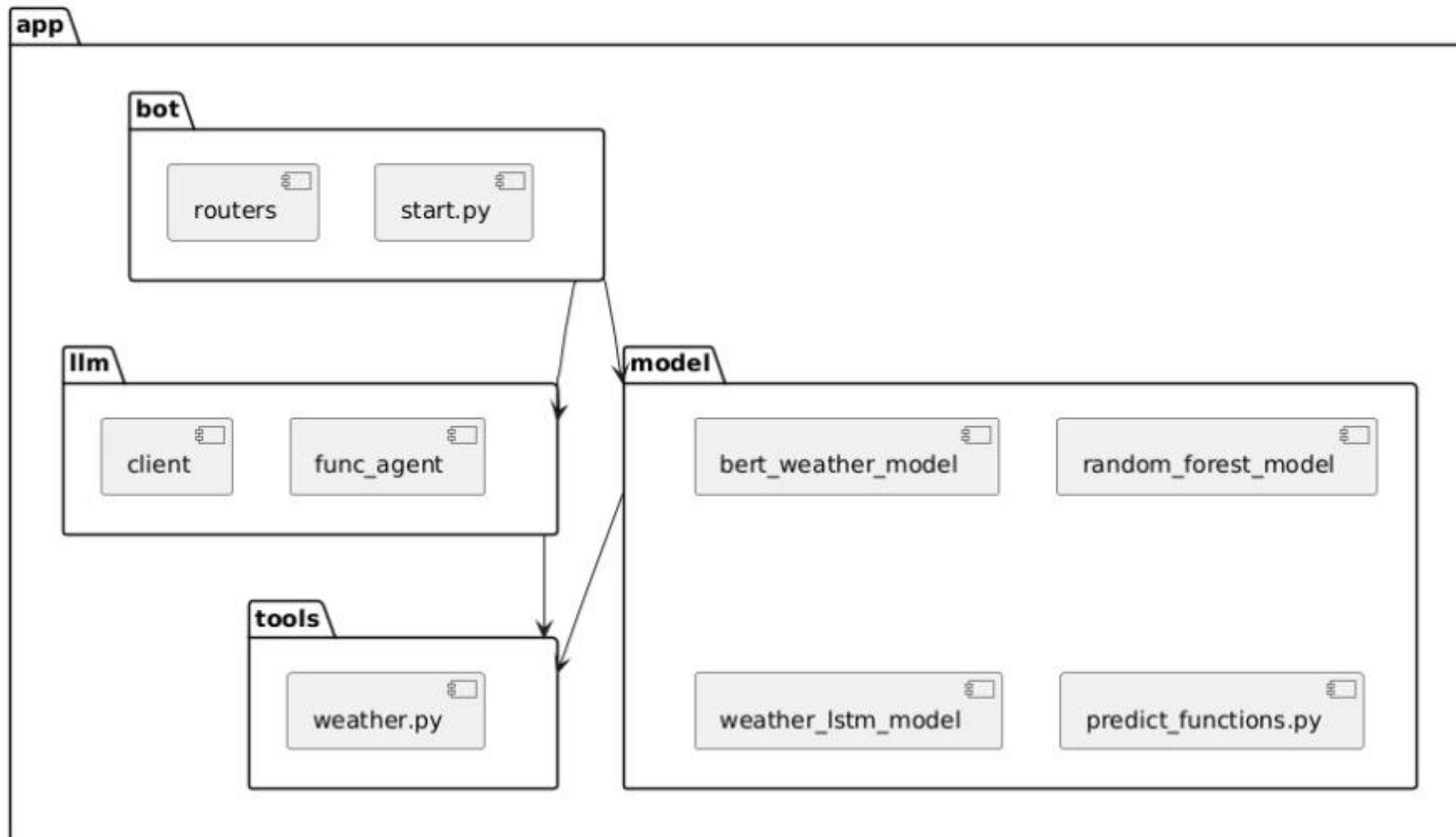
Вспомогательные компоненты

- Tools
 - Утилиты для обработки данных
 - Функции работы с погодными данными
- Client
 - Интерфейс взаимодействия с API
- Func_agent
 - Агенты для обработки запросов

Архитектура системы



Архитектура системы





Демонстрация



LSTM

Через час, в 17:03, ожидается температура около -0.2°C .
Рекомендуется надеть теплую куртку, шапку, перчатки и шарф.

16:03

Какая погода будет через час?

16:03 ✓✓

Через пять часов, в 21:04, ожидается температура около -0.2°C .
Рекомендуется надеть теплую куртку, шапку, шарф и перчатки, чтобы не замерзнуть.

16:04

Какая погода будет через 5 часов?

16:04 ✓✓

BERT

какая погода будет завтра? 17:59 ✓✓

Завтра, 28 декабря 2024 года, в 17:59 ожидается морозная погода со средней температурой около -2 градусов Цельсия. Количество осадков составит примерно 0,6 мм, а облачность достигнет 91%, так что одевайтесь теплее! Рекомендуется надеть теплый свитер или кофту под куртку, а также шапку, перчатки и шарф для защиты от холода.

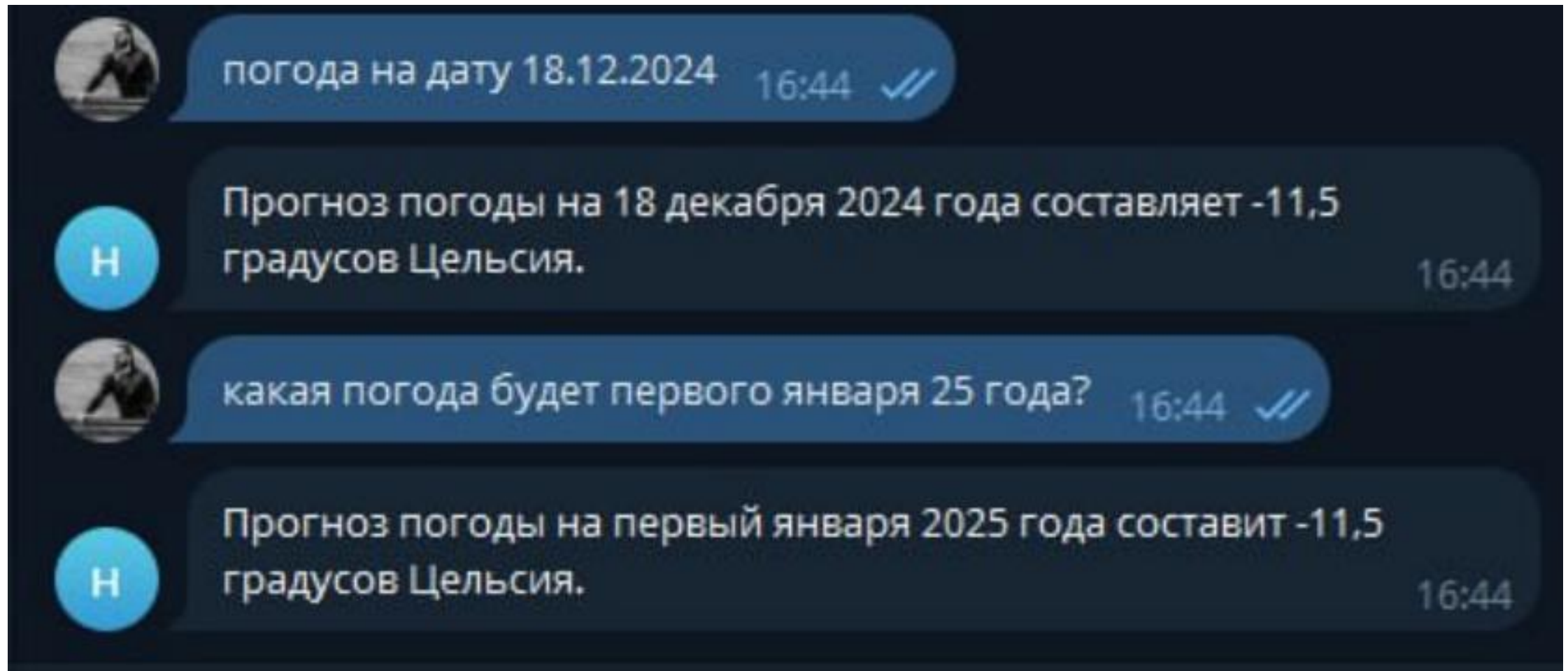
17:59

какая погода будет через 2 недели? 18:03 ✓✓

Через две недели, 13 января 2025 года, ожидается следующая погода: температура воздуха составит около -1.29 градусов Цельсия. Прогнозируется небольшой снег с количеством осадков 0.91 мм. Облачность будет почти полной — 95.04%.

18:03

Random Forest





Планы развития



Планы развития

- Интеграция компьютерного зрения:
 - Внедрение YOLO 11 для классификации одежды
 - Распознавание типов одежды по фотографиям пользователей
 - Создание базы данных одежды с характеристиками

Планы развития

- Разработка RAG системы:
 - Создание базы знаний о соответствии одежды погодным условиям
 - Интеграция с существующей системой прогнозирования
 - Разработка алгоритмов подбора одежды на основе:
 - Прогноза температуры
 - Осадков
 - Силы ветра
 - Времени года

Планы развития

- Улучшение пользовательского опыта:
 - Добавление функции загрузки фотографий в Telegram бот
 - Персонализированные рекомендации по гардеробу
 - Интерактивный интерфейс для управления базой одежды



Спасибо за внимание!