Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2024, Том 16, № 3 / 2024, Vol. 16, Iss. 3 https://esj.today/issue-3-2024.html

URL статьи: https://esj.today/PDF/07NZVN324.pdf

1.6.21. Геоэкология (технические науки)

Ссылка для цитирования этой статьи:
Жукова, Ж. С. Сравнение методов прогнозирования температур по данным штата Квинсленд, Австралия / Ж. С. Жукова, В. В. Ерофеева, Т. А. Тагоев, Н. Е. Убушаев, И. А. Фёдоров // Вестник евразийской науки. — 2024. — Т. 16. — № 3. — URL: https://esj.today/PDF/07NZVN324.pdf

For citation:
Zhukova Zh.S., Erofeeva V.V., Tagoyev T.A., Ubushayev N.Ye., Fyodorov I.A. Comparison of temperature forecasting methods based on data from the State of Queensland, Australia. *The Eurasian Scientific Journal*. 2024;16(3): 07NZVN324. Available at: https://esj.today/PDF/07NZVN324.pdf. (In Russ., abstract in Eng.)

УДК 004.942; 551.583.1

Жукова Жанна Сергеевна
Ордена Трудового Красного Знамени
ФГБОУ ВО «Московский технический университет связи и информатики», Москва, Россия
Старший преподаватель
E-mail: assamblee@mail.ru
ORCID: https://orcid.org/0000-0002-9828-4413
РИНЦ: https://elibrary.ru/author_profile.asp?id=1061900
SCOPUS: https://www.scopus.com/authid/detail.url?authorId=57219179989

Ерофеева Виктория Вячеславовна
Ордена Трудового Красного Знамени
ФГБОУ ВО «Московский технический университет связи и информатики», Москва, Россия
ФГАОУ ВО «Российский университет дружбы народов имени Патриса Лумумбы», Москва, Россия
Доцент
Кандидат биологических наук, доцент
E-mail: erofeeva-viktori@mail.ru
ORCID: https://orcid.org/0000-0002-0236-1876

Тагоев Тимур Ахмаджонович
Ордена Трудового Красного Знамени
ФГБОУ ВО «Московский технический университет связи и информатики», Москва, Россия
E-mail: mifers@mail.ru

Убушаев Никита Евгеньевич
Ордена Трудового Красного Знамени
ФГБОУ ВО «Московский технический университет связи и информатики», Москва, Россия
E-mail: nikita.ubushaec@gmail.ru

Фёдоров Игорь Александрович
Ордена Трудового Красного Знамени
ФГБОУ ВО «Московский технический университет связи и информатики», Москва, Россия
E-mail: powerfed@mail.ru

Сравнение методов прогнозирования температур по данным штата Квинсленд, Австралия

Аннотация. В работе авторами представлены результаты прогнозирования температурных изменений на основе массива данных пяти метеорологических станций штата Квинсленд, Австралия. Было проведено сравнение методов random forest regressor, k-nearest

neighbors, linear regression, seasonal autoregressive integrated moving average для прогнозирования изменения температур до 2030 года на разных широтах штата с использованием методов регрессии и метода временных рядов seasonal autoregressive integrated moving average. Фактические значения температуры, используемые для прогнозирования, были собраны в два файла: «kvin2.xlsx» и «kvshir1.xlsx». «kvin2.xlsx» содержит фактические температуры для пяти станций за период с 1856 по 2022 год, в то время как «kvshir1.xlsx» содержит фактические температуры для 236 станций за тот же период. Более подробное сравнение изменчивости температуры из файла «kvin2.xlsx» было проведено с использованием метода случайного леса для пяти станций. Кроме того, точность прогноза для этих станций была рассчитана для двух прогонов, поскольку при прогнозировании использовался случайный разброс с использованием метода случайного леса. Каждый прогон программы выдает новые значения, основанные на тех, которые доступны в файле «kvin2.xlsx». В итоге были получены предсказания, учитывающие случайные величины, различные (но не значительно) для каждого прогона программы. Точность была рассчитана путем сравнения прогнозируемых температур для двух прогонов с фактическими температурами из первого файла. Для прогнозирования температуры из файла «kvshir1.xlsx» применялись такие методы, как k-nearest neighbors, линейная регрессия и seasonal autoregressive integrated moving average, без использования случайного разброса. Для этого из множества станций было выбрано девять с длинными рядами наблюдений, т. к. этого количества достаточно для наглядной демонстрации работы программы. В результате сравнения различных методов при прогнозировании random forest regressor показал, что данный метод прогнозирует значения температур с точностью не ниже 96,164 %, самая маленькая среднеквадратичная ошибка высчитывается в методе k-nearest neighbors: 0,175. На основе random forest regressor было проведено прогнозирование по пяти станциям до 2030 года.

Ключевые слова: Австралия; климат; прогнозирование; random forest regressor; k-nearest neighbors; linear regression; seasonal autoregressive integrated moving average

Введение

Климат на планете подвержен постоянным колебаниям, оказывает прямое воздействие на эволюционное развитие человеческого общества [1]. Согласно исследованиям за последние 420 тысяч лет было 4 периода похолодания, которые чередовались с периодами межледниковья. Современный период голоцен длится порядка 12 тысяч лет, характеризуется комфортными условиями существования, малым перемещением континентов и средней температурой 14°C. Он делится на три периода, за время которых было 13 взаимносменяющихся периодов потепления и похолодания. На сегодняшний момент изучены керны льда Гренландии и Антарктиды (озеро Восток) и по ним восстановлен климат планеты на несколько сот тысячелетий назад [2; 3]. Около 5,5 тысяч лет назад наступил голоценовый оптимум, на который пришёлся расцвет многих цивилизаций. В дальнейшем произошло снижение температуры и влажности воздуха, которое продлилось около 2,5 тысяч лет. С середины XIX века началось потепление, которое связывают с увеличением концентрации CO_2 (в том числе техногенного характера) и антропогенным давлением на биосферу [4; 5]. Таким образом изучение климата является актуальным направлением современной науки. Прогнозирование изменений климата важно не только для оценки состояния окружающей среды, но и для устойчивого экономического развития, благополучия человеческой цивилизации, в основе которого лежит сельское хозяйство [1].

Современное потепление климата планеты заметно даже в рамках наблюдений одного поколения. Незначительные колебания не оказывают заметного влияния на развитие техносферы как в краткосрочной, так и в долгосрочной перспективе. Однако значительное

изменение температуры планеты может привести к изменениям в качестве и количестве потребляемых ресурсов, их доступности, что может повлиять на выживаемость человеческой популяции [6]. Анализируя температурные изменения прошлого, можно с использованием современных информационных технологий строить прогнозы на будущее.

Возможности искусственного интеллекта позволяют разрабатывать новые подходы к изучению состояния климата. Происходит переход от традиционных методов, направленных на понимание закономерностей, к предсказаниям по большим массивам данных, которые регулярно обновляются [7; 8].

Методы искусственного интеллекта и машинное обучение являются ключевыми при работе с возрастающими объёмами климатических данных. В климатологии они используются не только для анализа, но показали свою эффективность при заполнении пробелов наблюдений [9; 10]. Концепция IoT позволяет объединять устройства и машины для сбора и обработки больших объёмов данных через интернет [9; 11].

Целью работы — сравнение методов обработки доступных массивов данных о температуре по нескольким метеорологическим станциям штата Квинсленд, Австралия за исторический период до 2018 года, проведение контрольного прогнозирования по следующим пяти годам с использованием искусственного интеллекта и итоговое предсказание изменения температур до 2030 года.

Материалы и методы

Для анализа и предсказания климатических изменений в данном исследовании был выбран штат Квинсленд, Австралия, расположенный на северо-востоке. На континенте много метеорологических станций, где проводятся регулярные наблюдения за климатом и имеются длинные температурные ряды. В работе использовались данные средних годовых температур по станциям.

Для работы с данными были созданы два файла kvshir1.xlsx, в котором собраны фактические температуры для 236 станций штата (за период наблюдений с 1856 года), и файл kvin2.xlsx с данными пяти исследуемых станций, выбранных в разных широтах.

Метод random forest regressor на основе файла «kvin2.xlsx»

Для сравнения максимальных прогнозируемых температур с минимальными прогнозируемыми температурами и прогнозируемых температур с реальными температурами был использован метод Random forest regressor (регрессор случайного леса). Случайный лес — алгоритм машинного обучения, который заключается в использовании ансамбля (совокупности) деревьев решений (decision trees), формула для регрессии [12]:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} b_i(x); \quad (1)$$

где $\hat{y}_i$ — предсказанное значение температуры; n — количество деревьев в ансамбле; $b_i(x)$ — прогнозируемое значение температуры, полученное от i-го дерева в ансамбле.

Для каждого места строится модель случайного леса. Обучение происходит на основе относительных значений годов (разница между годом и первым годом в данных) и реальных температурных значений, из каждого года вычитается первый год в DataFrame (наши данные). В этом коде использовался случайный разброс.

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
# Загрузка данных из файла kvin2.xlsx
temp_df = pd.read_excel('kvin2.xlsx')
# Заменяем отсутствующие значения на среднее значение для каждого
места
temp_df.fillna(temp_df.mean(), inplace = True)
# Обучение модели случайного леса для каждого места
models = {}
for place in temp_df.columns[1:]:  # Исключаем столбец с годом
    model = RandomForestRegressor()
    X = temp_df['год'].values.reshape(-1, 1) —
temp_df['год'].values[0]  # Преобразуем годы в относительные
значения
    y = temp_df[place].values
    model.fit(X, y)
    models[place] = model
# Прогнозирование температурных значений с 2018 по 2030 год
predtemps = {}
future_years = np.arange(2018, 2031).reshape(-1, 1)
for place, model in models.items():
    predtemps[place] = model.predict(
        future_years — temp_df['год'].values[0])  # Прогноз на
основе относительных значений
    # Расчет стандартного отклонения прогнозов
    std_dev = np.std(predtemps[place])
    # Добавление разброса для годов после 2022
    for i, year in enumerate(range(2023, 2031)):
        predtemps[place][i + 5] += std_dev * np.random.randn()
# Добавляем случайную составляющую
# Сравнение реальной и прогнозируемой температуры для каждого места
на одном графике
plt.figure(figsize =(20, 10))
for i, place in enumerate(temp_df.columns[1:], start = 1):
    plt.subplot(2, len(temp_df.columns[1:]), i)
    plt.bar(['Реальная', 'Прогнозируемая'], [temp_df[place].iloc
[-1], predtemps[place][-1]], width = 0.4, color = ['red', 'blue'],
alpha = 0.5)
    if i == 1:  # Добавляем подпись только на первый график
        plt.ylabel('Температура, °C')
    plt.title(place)
# Сравнение максимальной и минимальной прогнозируемой температуры
для каждого места
plt.figure(figsize = (20, 10))
for i, place in enumerate(temp_df.columns[1:], start = 1):
    plt.subplot(2, len(temp_df.columns[1:]), i)
    plt.bar(['Максимальная', 'Минимальная'],
[max(predtemps[place]), min(predtemps[place])], width = 0.4,
color = ['red', 'blue'], alpha = 0.5)
```


07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

```
    if i == 1:  # Добавляем подпись только на первый график
        plt.ylabel('Температура, °C')
    plt.title(place)
# Построение графика изменения регрессии для каждого места
plt.figure(figsize=(10, 6))
for place, model in models.items():
    X_pred = future_years - temp_df['год'].values[0]
    y_pred = model.predict(X_pred)
    # Построение графика прогнозируемых температур для каждого
места
    plt.plot(np.arange(2018, 2031), predtemps[place])
# Добавляем аннотации с названиями мест рядом с кривыми на графике
    plt.annotate(place, xy = (2030, predtemps[place][-1]),
xytext = (5, -5), textcoords = 'offset points', fontsize = 8)
    # Добавляем текст с прогнозируемой температурой для каждого
года после 2022, отмеченного на оси x
    for year, temp in zip(range(2023, 2031), predtemps[place][5:]):
        if year in plt.xticks()[0]:  # Проверяем, что год
присутствует на оси x
            plt.text(year, temp, f'{temp:.2f}', ha = 'right',
va = 'bottom', fontsize = 8)
plt.title('Прогнозируемые температуры для каждого места с 2018 по
2030 год')
plt.xlabel('Год')
plt.ylabel('Температура, °C')
plt.legend()
plt.grid(True)
plt.show()
```

Листинг кода 1: *random forest (составлено авторами)*

```
import numpy as np
import pandas as pd
# Функция для вычисления точности в процентах
def calculate_accuracy(actual, predicted):
    accuracy = []
    for i in range(len(actual)):
        actual_diff = actual[i] - predicted[i]
        accuracy.append((1 - abs(actual_diff) / actual[i]) * 100)
    return accuracy
# Загрузка данных из файла
temp_df = pd.read_excel('kvin2.xlsx')
 #Заменяем отсутствующие значения на среднее значение для каждого
места
temp_df.fillna(temp_df.median(), inplace=True)
# Преобразуем годы в относительные значения
temp_df['год'] = temp_df['год'].values - temp_df['год'].values[0]
# Предсказанные температуры для каждой станции при первом и втором
запусках
predicted_temperatures = {
    'Вейпа': [(26.86, 26.85), (26.86, 26.85)],
    'Аэропорт Локхарт': [(26.14, 26.01), (26.14, 26.01)],
```


07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

```
    'Арчерфилд': [(20.00, 19.84), (20.00, 19.84)],
    'Амберли': [(19.00, 19.20), (19.00, 19.20)],
    'Аплторп': [(14.04, 14.42), (14.04, 14.42)]
}
# Вычисление точности для каждой станции и каждого запуска
for place in temp_df.columns[1:]:  # Исключаем столбец с годом
    print(f"Станция: {place}")
    actual_median = np.median(temp_df[place].values)
    print(f"Медиана фактических температур: {actual_median}")
    for i, (predicted_temp_1, predicted_temp_2) in
enumerate(predicted_temperatures[place], 1):
        predicted_temps = [predicted_temp_1, predicted_temp_2]
        accuracy = calculate_accuracy([actual_median],
predicted_temps)
        print(f"Точность для запуска {i}: {accuracy}")
```

Листинг кода 2: *вычисление точности прогнозирования для каждой станции первого файла при первом и втором запусках программы методом random forest regressor (составлено авторами)*

В каждом из ниже рассмотренных методов будут прогнозироваться значения для первых девяти станций (количество станций достаточное для наглядного представления о работе методов) без использования случайного разброса, так как вычислялась средняя квадратичная ошибка (MSE).

Метод k-nearest neighbors (KNN) на основе файла «kvshir1.xlsx»

Метод k ближайших соседей (KNN) используется для решения задачи регрессии. Сначала данные разбиваются на обучающий и тестовый наборы. Далее модель KNN обучается на обучающих данных, где для каждого объекта хранится его признаковое описание и соответствующее значение целевой температуры. Для предсказания значений на тестовом наборе температур KNN ищет k ближайших соседей к каждому объекту и использует их для вычисления предсказанного значения. Затем оценивается качество модели с помощью среднеквадратичной ошибки (MSE), которая измеряет разницу между реальными и предсказанными значениями:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2; \quad (2)$$

где n — количество температур на станции; y_i — фактическое значение зависимой переменной; $\hat{y}_i$ — предсказанное значение зависимой переменной [12].

В нашем случае используется метрика среднего арифметического температур (предсказанные значения примерные) для каждой станции, но также существуют и другие популярные метрики для KNN, например: евклидово расстояние, косинусное расстояние и манхэттенское расстояние. Т. к. в коде не указано количество соседей, которые будут использованы для прогнозирования, для функции «train_test_split» по умолчанию оно равно 5.

```
import numpy as np
import pandas as pd
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# Загрузка данных из файла с корректным чтением заголовков и
```


07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

```
индексов, пропуская первые три строки
temp_df = pd.read_excel('kvshir1.xlsx', header = 2, index_col = -1)
# Заполнение отсутствующих значений средними
temp_df.fillna(temp_df.mean(), inplace = True)
# Удаляем первый столбец с годом
temp_df = temp_df.iloc[:, 1:]
# Определение признаков (X) и целевой переменной (y)
X = temp_df.iloc[:, 1:]  # Признаки
y = temp_df.iloc[:, 0]   # Целевая переменная
# Преобразование имен столбцов в строковый тип данных
X.columns = X.columns.astype(str)
# Разделение данных на обучающий и тестовый наборы
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state = 42)
# Получение названий станций из третьей строки
stations = temp_df.columns[0:9]
# Обучение модели KNeighborsRegressor
knn_model = KNeighborsRegressor()
knn_model.fit(X_train, y_train)
# Предсказание для тестового набора данных и оценка качества модели
print("Модель: K Nearest Neighbors")
predictions = knn_model.predict(X_test)
mse = mean_squared_error(y_test, predictions)
print(f"Mean Squared Error (K Nearest Neighbors): {mse}")
for station, prediction in zip(stations, predictions):
    print(f"Станция: {station}, Предсказанное значение:
{prediction}")
```

Листинг кода 3: *KNN (составлено авторами [13–15])*

Метод Linear Regression на основе файла «kvshir1.xlsx»

Линейная регрессия — метод статистического моделирования, который используется для изучения отношений между зависимой переменной (целевой) и одной или несколькими независимыми переменными (признаками). В коде данные разделяются на обучающий и тестовый наборы, чтобы оценить качество модели [12]. В библиотеке scikit-learn коэффициенты линейной регрессии инициализируются случайным образом перед обучением модели. Это означает, что по умолчанию их значения не заданы заранее и зависят от начального состояния генератора случайных чисел. Каждый раз при обучении модели они могут быть разными. Таким образом, хотя начальные значения коэффициентов могут быть случайными, после обучения они фиксируются и остаются постоянными для последующих прогнозов температур. Данная модель линейной регрессии обучается на обучающих данных, где она подстраивает коэффициенты линейной комбинации признаков для предсказания температур на станциях. Затем линейная регрессия применяется к тестовым данным, и для каждой станции вычисляются прогнозные значения температуры. Прогнозы модели основаны на оцененных коэффициентах, которые настраиваются таким образом, чтобы минимизировать сумму квадратов разностей между реальными и предсказанными значениями (SSE), так как для линейной регрессии $\hat{y}_i = a + b * x_i$, то:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - b * x_i)^2; \quad (3)$$

где y_i — фактическое значение зависимой переменной; $\hat{y}_i$ — предсказанное значение

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

зависимой переменной; a — коэффициент сдвига; b — коэффициент наклона; x_i — независимая переменная [12].

Среднеквадратичная ошибка (MSE) используется для оценки качества модели. Таким образом, линейная регрессия пытается найти линейную зависимость между признаками и температурой на станциях.

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# Загрузка данных из файла с корректным чтением заголовков и
индексов, пропуская первые три строки
temp_df = pd.read_excel('kvshir1.xlsx', header = 2, index_col = -1)
# Заполнение отсутствующих значений средними
temp_df.fillna(temp_df.mean(), inplace = True)
# Удаляем первый столбец с годом
temp_df = temp_df.iloc[:, 1:]
# Определение признаков (X) и целевой переменной (y)
X = temp_df.iloc[:, 1:]  # Признаки
y = temp_df.iloc[:, 0]   # Целевая переменная
# Преобразование имен столбцов в строковый тип данных
X.columns = X.columns.astype(str)
# Разделение данных на обучающий и тестовый наборы
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = 0.2, random_state = 42)
# Получение названий станций из третьей строки
stations = temp_df.columns[0:9]
# Обучение модели LinearRegression
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
# Предсказание для тестового набора данных и оценка качества модели
print("\nМодель: Linear Regression")
predictions = lr_model.predict(X_test)
mse = mean_squared_error(y_test, predictions)
print(f"Mean Squared Error (Linear Regression): {mse}")
for station, prediction in zip(stations, predictions):
    print(f"Станция: {station}, Предсказанное значение:
{prediction}")
```

Листинг кода 4: *linear resgression (составлено авторами [13–16])*

Метод SARIMA на основе файла «kvshir1.xlsx»

Метод seasonal autoregressive integrated moving average (SARIMA) используется для анализа и прогнозирования временных рядов с учетом сезонных изменений, трендов и структуры данных. Данный метод использует модель авторегрессии интегрированного скользящего среднего (ARIMA), которая представляет собой комбинацию компонентов авторегрессии (AR), интегрирования (I) и скользящего среднего (MA). В коде модель SARIMA применяется к временным рядам температурных данных на различных станциях:

$$y_t = c + \varphi_1 y_{t-1} + .. + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + .. + \theta_q \varepsilon_{t-q} + \Theta_1 \varepsilon_{t-s} + .. + \Theta_s \varepsilon_{t-s} + \varepsilon_t; \qquad (4)$$

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

где y_t — текущее значение временного ряда; c— константа; φ_i — коэффициенты авторегрессии; p — порядок авторегрессии; q — порядок скользящего среднего; ε_t — случайная ошибка в момент времени t; θ_i — коэффициенты скользящего среднего; s — период сезонности; Θ_i — коэффициенты сезонного скользящего среднего.

Модель SARIMA определяется параметрами order = (1, 1, 1) для авторегрессии, дифференциации и скользящего среднего, а также seasonal_order = (1, 1, 1, 12) для учета сезонности с периодом в 12 месяцев. После обучения модели SARIMA на обучающем наборе она используется для прогнозирования значений температуры на тестовом наборе данных. Полученные прогнозы сравниваются с фактическими значениями, и их качество оценивается с помощью среднеквадратичной ошибки (MSE). Прогнозы модели основаны на уравнении временного ряда, в котором учитываются лаги временного ряда, сезонные компоненты и ошибка модели.

```
import numpy as np
import pandas as pd
from statsmodels.tsa.arima.model import ARIMA
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# Загрузка данных из файла с корректным чтением заголовков и
индексов, пропуская первые три строки
temp_df = pd.read_excel('kvshir1.xlsx', header = 2, index_col = -1)
# Заполнение отсутствующих значений средними
temp_df.fillna(temp_df.mean(), inplace = True)
# Удаляем первый столбец с годом
temp_df = temp_df.iloc[:, 1:]
# Определение признаков (X) и целевой переменной (y)
X = temp_df.iloc[:, 1:]  # Признаки
y = temp_df.iloc[:, 0]   # Целевая переменная
# Преобразование имен столбцов в строковый тип данных
X.columns = X.columns.astype(str)
# Разделение данных на обучающий и тестовый наборы
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
= 0.2, random_state = 42)
# Получение названий станций из третьей строки
stations = temp_df.columns[0:9]
# Обучение модели SARIMA
sarima_model = ARIMA(y_train, order = (1, 1, 1),
seasonal_order = (1, 1, 1, 12))
sarima_model_fit = sarima_model.fit()
# Предсказание для тестового набора данных и оценка качества модели
print("\nМодель: SARIMA")
predictions = sarima_model_fit.forecast(steps=len(X_test))
mse = mean_squared_error(y_test, predictions)
print(f"Mean Squared Error (SARIMA): {mse}")
for station, prediction in zip(stations, predictions):
    print(f"Станция: {station}, Предсказанное значение:
{prediction}")
```

Листинг кода 5: *SARIMA (составлено авторами [13; 14; 17])*

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Результаты и обсуждения

Результаты работы для первого файла для сравнения максимальных и минимальных прогнозируемых температур, а также реальных и прогнозируемых температур для пяти станций при первом и втором запусках представлены на рисунках 1, 2:

Результат листинга кода 1

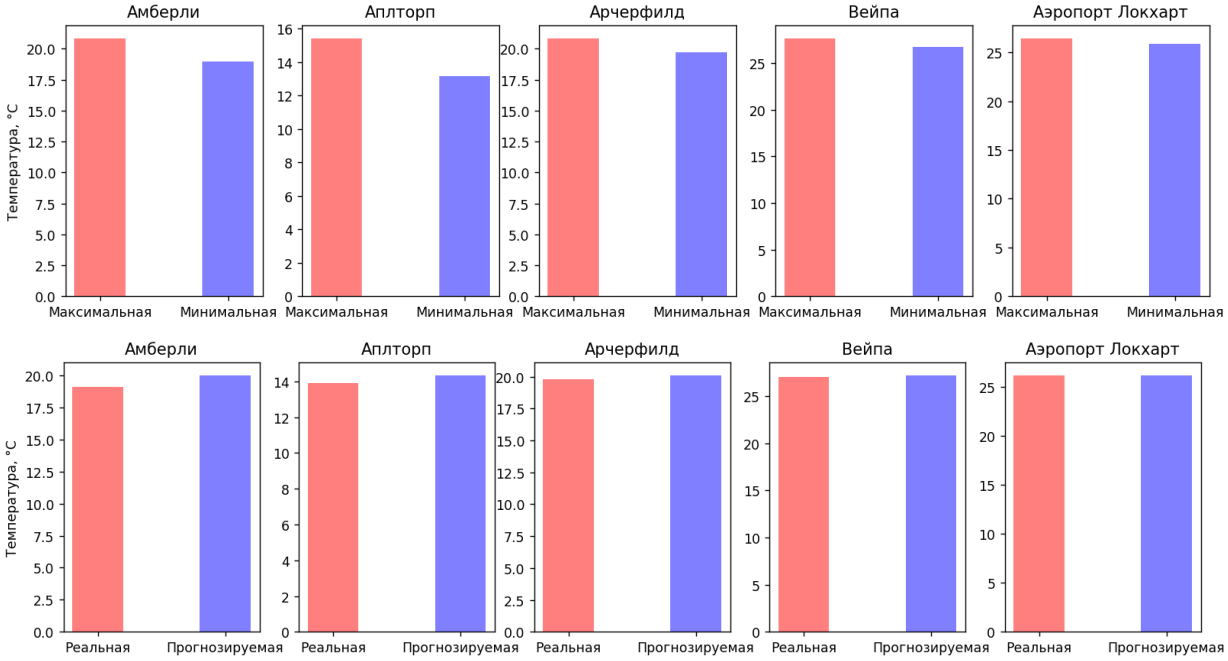

Рисунок 1. *Диаграммы сравнения максимальных и минимальных прогнозируемых температур, а также реальных и прогнозируемых температур для пяти станций при первом запуске (составлено авторами)*

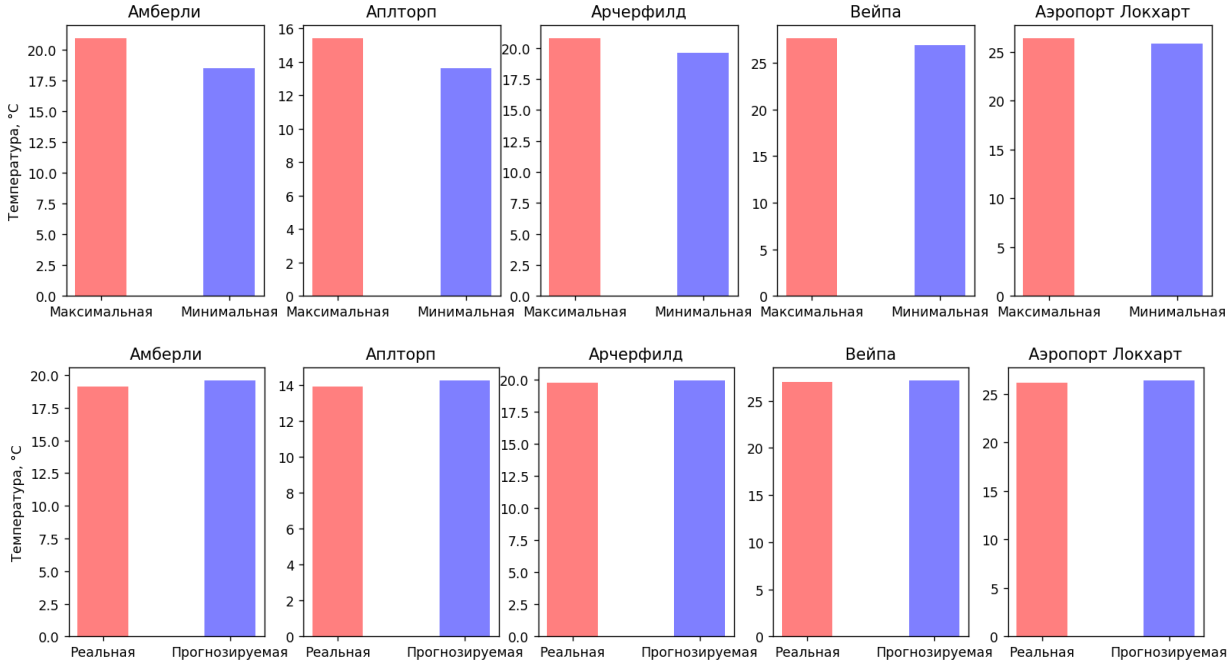

Рисунок 2. *Диаграммы сравнения максимальных и минимальных прогнозируемых температур, а также реальных и прогнозируемых температур для пяти станций при втором запуске (составлено авторами)*

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Как видно из графиков, каждый раз при запуске программы, прогнозируются новые значения для максимальной и минимальной прогнозируемых температур, а также для прогнозируемых температур при сравнении с реальными температурами.

Результаты работы программы для первого файла для вычисления точности прогнозирования для каждой станции при первом и втором запусках программы метода random forest regressor, также здесь, как упоминалось ранее, используется случайный разброс (рис. 3, 4):

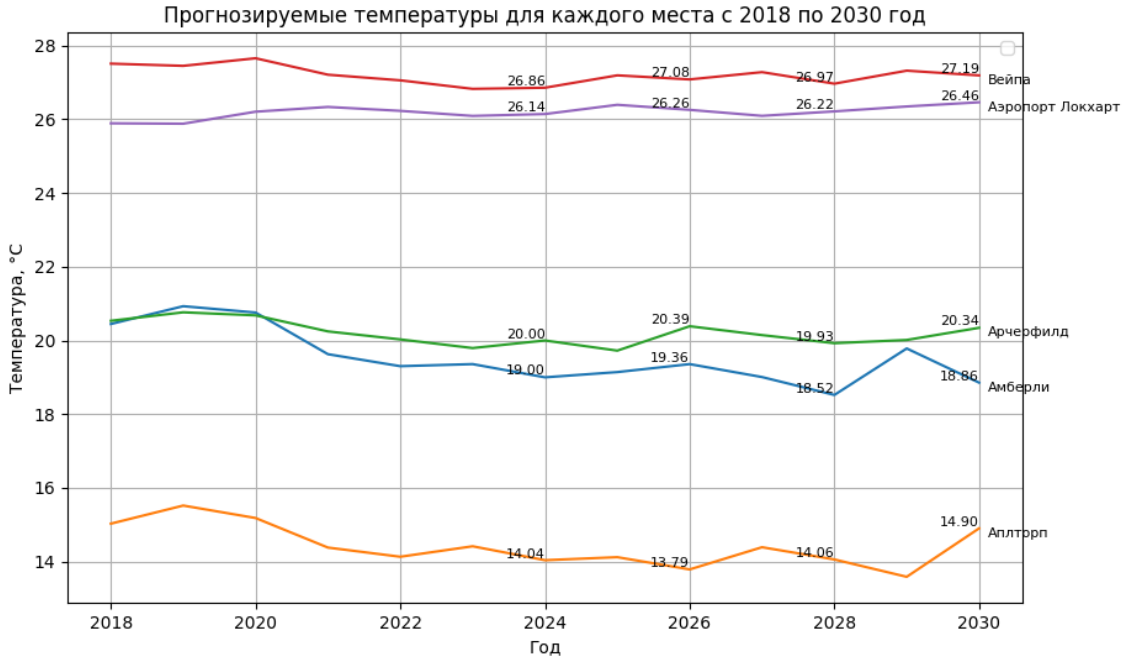

Рисунок 3. *График прогнозируемых значений после 2022 года при первом запуске программы (составлено авторами)*

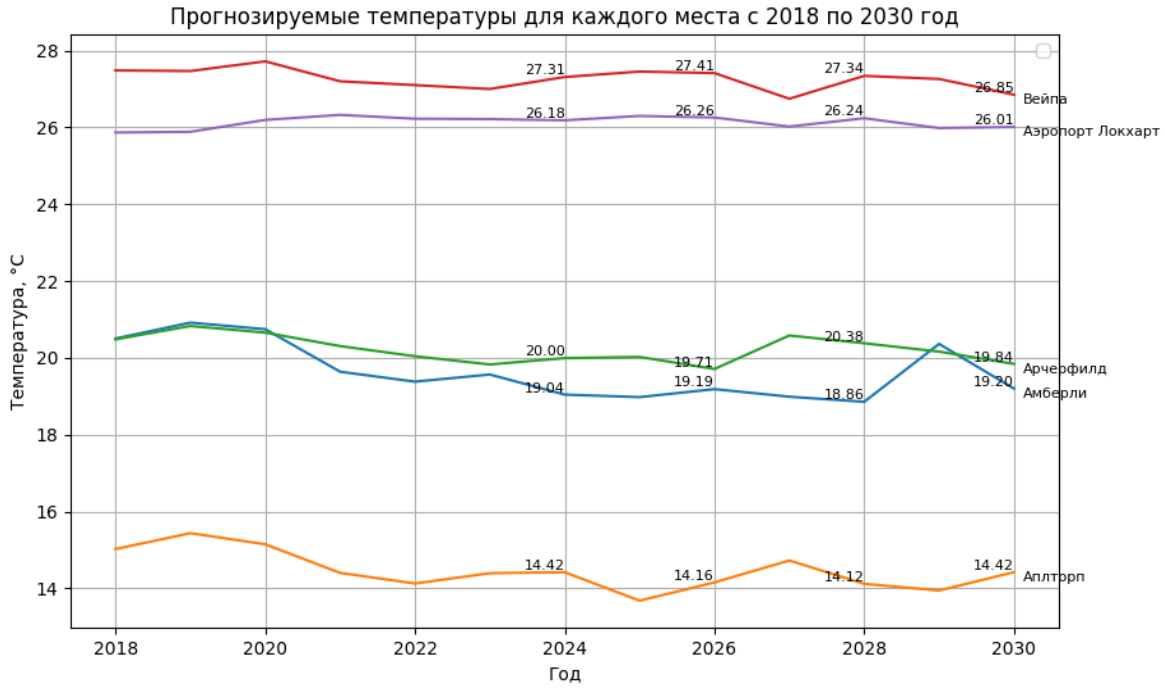

Рисунок 4. *График прогнозируемых значений после 2022 года при втором запуске программы (составлено авторами)*

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Результат листинга кода 2

Точность прогнозирования значений после 2022 года на основе медиан фактических значений для каждой станции первого файла при двух запусках методом random forest regressor в процентах составляют соответственно:

Медиана фактических температур: 19.3

Точность для запуска 1: [98.44559585492227]

Точность для запуска 2: [98.44559585492227]

Станция: Аплторп

Медиана фактических температур: 14.6

Точность для запуска 1: [96.16438356164383]

Точность для запуска 2: [96.16438356164383]

Станция: Арчерфилд

Медиана фактических температур: 20.1

Точность для запуска 1: [99.50248756218905]

Точность для запуска 2: [99.50248756218905]

Станция: Вейпа

Медиана фактических температур: 26.8

Точность для запуска 1: [99.77611940298507]

Точность для запуска 2: [99.77611940298507]

Станция: Аэропорт Локхарт

Медиана фактических температур: 25.8

Точность для запуска 1: [98.68217054263566]

Точность для запуска 2: [98.68217054263566]

Результат работы программы для второго файла для прогнозирования температур на девяти станциях и оценки программы путем вычисления средней квадратичной ошибки (MSE) методом KNN.

Результат листинга кода 3

Model: K Nearest Neighbors

Mean Squared Error (K Nearest Neighbors): 0.17488977176669432

Station: Adavale Post Office, Predicted value: 21.086813186813185

Station: Alva Beach, Predicted value: 21.086813186813185

Station: Beerwah-Crohamhurst, Predicted value: 21.086813186813185

Station: Biloela Dpi, Predicted value: 21.086813186813185

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Station: Birdsville Police Station, Predicted value: 21.086813186813185

Station: Bottom Of Pin Gin Hill, Predicted value: 21.086813186813185

Station: Bowen Post Office, Predicted value: 21.086813186813185

Station: Brigalow Research Stn, Predicted value: 21.086813186813185

Station: Brisbane Regional Office, Predicted value: 21.774725274725277

Результат работы программы для второго файла для прогнозирования температур на девяти станциях и оценки программы путем вычисления средней квадратичной ошибки (MSE) методом linear regression.

Результат листинга кода 4

Model: Linear Regression

Mean Squared Error (Linear Regression): 0.3245345199677637

Station: Adavale Post Office, Predicted value: 21.336593938696538

Station: Alva Beach, Predicted value: 21.08681318681316

Station: Beerwah- Crohamhurst, Predicted value: 21.471082935855996

Station: Biloela Dpi, Predicted value: 21.08681318681316

Station: Birdsville Police Station, Predicted value: 21.380457652100112

Station: Bottom Of Pin Gin Hill, Predicted value: 20.789295296562102

Station: Bowen Post Office, Predicted value: 21.514922694821152

Station: Brigalow Research Stn, Predicted value: 20.987318967308674

Station: Brisbane Regional Office, Predicted value: 23.224015508828465

Результат работы программы для второго файла прогнозирования температур на девяти станциях и оценки программы путем вычисления средней квадратичной ошибки (MSE) методом SARIMA.

Результат листинга кода 5

Model: SARIMA

Mean Squared Error (SARIMA): 28.137874408759814

Station: Adavale Post Office, Predicted value: 20.66434545987257

Station: Alva Beach, Predicted value: 20.663234835457686

Station: Beerwah- Crohamhurst, Predicted value: 21.24987888300953

Station: Biloela Dpi, Predicted value: 20.67123674407488

Station: Birdsville Police Station, Predicted value: 20.663427604359878

Station: Bottom Of Pin Gin Hill, Predicted value: 4.1407780382332255

07NZVN324
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Station: Bowen Post Office, Predicted value: 20.664284254002098

Station: Brigalow Research Stn, Predicted value: 20.66323456743976

Station: Brisbane Regional Office, Predicted value: 20.663233345194705

Точность прогнозов для каждой модели, как правило, зависит от различных факторов, таких как размер исходного набора данных и количество входных параметров (гиперпараметров). К ним относятся глубина деревьев в случайном лесу, скорость обучения при градиентном ускорении, коэффициент регуляризации в линейных моделях, количество соседей в методе k ближайших соседей и различные показатели, используемые для оценки модели. Детальное сравнение различных методов искусственного интеллекта для прогнозирования значений температуры остается актуальным, поскольку каждый метод предсказывает разные значения.

Сравнение температур на пяти станциях для первого файла методом регрессии случайного леса показало, что самые большие максимальные и минимальные температуры прогнозируются на станциях Вейпа и Аэропорт Локхарт, а самые маленькие на станциях Амберли и Аплторп. Точность прогнозирования на основе медиан фактических значений для каждой станции первого файла при двух запусках для каждой из пяти станций методом random forest regressor показала, что данный метод прогнозирует значения температур с точностью не ниже 96 % при двух запусках, самый точный прогноз здесь получился для станции Вейпа — 99,776 %. Сравнение трех методов для второго файла показало, что самая маленькая среднеквадратичная ошибка (MSE) высчитывается в методе KNN: 0,175. И значит, именно KNN является самым точным методом из трех. В методе линейной регрессии видна самая большая максимально предсказанная температура, значит именно линейная регрессия спрогнозировала самую большую среднюю температуру среди трех методов: 21,216. А самую маленькую среднюю температуру среди трех методов спрогнозировал метод SARIMA: 19,006. Прогнозирование температур до 2030 года по пяти станциям штата Квинсленд было проведено по методу random forest regressor.

ЛИТЕРАТУРА

1. Жукова, Ж.С. Исследование вариативности температурных показателей Антарктиды / Ж.С. Жукова, В.В. Ерофеева // Вопросы науки. — 2023. — № 3. — С. 53–57. — EDN WRUQZV.
2. Weninger, B. et al. Climate forcing due to the 8200 cal yr BP event observed at Early Neolithic sites in the eastern Mediterranean. Quaternary Research 66, 401–420 (2006).
3. Жукова, Ж.С. Проблемы глобального изменения климата / Ж.С. Жукова // Тенденции развития науки и образования. — 2024. — № 105-13. — С. 144–147. — DOI 10.18411/trnio-01-2024-661. — EDN AQZEJN.
4. Котляков В.М. О причинах и следствиях современных изменений климата // Солнечно-земная физика. 2012. Вып. 21. С. 110–114.
5. Умнякова, Н.П. Изменение климата и содержание загрязняющих веществ в атмосфере / Н.П. Умнякова, В.А. Смирнов // Биосферная совместимость: человек, регион, технологии. — 2021. — № 2(34). — С. 34–51. — DOI 10.21869/2311-1518-2021-34-2-34-51. — EDN HSJNDE.

6. Использование нейронных сетей для моделирования климатических изменений / Ж.С. Жукова, В.В. Ерофеева, А.А. Тимофеев-Каракозов, С.Л. Яблочников // Вестник Кыргызско-Российского Славянского университета. — 2024. — Т. 24, № 4. — С. 183–188. — DOI 10.36979/1694-500X-2024-24-4-183-188. — EDN SZVYKX.

7. Чуйков, Р.Я. Использование искусственного интеллекта и геоинформационных систем в изучении проблем глобальных изменений климата (обзор) / Р.Я. Чуйков, Ю.С. Чуйков // Астраханский вестник экологического образования. — 2023. — № 4(76). — С. 96–111. — DOI 10.36698/2304-5957-2023-4-96-111.

8. Dawson A.G. Ice Age Earth: Late Quaternary geology and climate // Routeledge Physical Environment series. L.; NY., 1992, p. 293.

9. Чуйков, Р.Я. Использование искусственного интеллекта и геоинформационных систем в изучении проблем глобальных изменений климата (обзор) / Р.Я. Чуйков, Ю.С. Чуйков // Астраханский вестник экологического образования. — 2023. — № 4(76). — С. 96–111. — DOI 10.36698/2304-5957-2023-4-96-111. — EDN BJJIBT.

10. Kadow C., Hall D.M., Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020; 13(7), 408–413.

11. Suchetha M., Anitha V., Al-Gaadi K.A. IoT based crop-field monitoring and irrigation automation. 2016 10th International Conference on Intelligent Systems and Control (ISCO). 2016. URL: https://sci-hub.ru/10.1109/ICISC.2018.8399118 (дата обращения: 15.04.2024).

12. Модель глубокого обучения RF «Случайный лес» для прогнозирования прибыли организации в условиях цифровой экономики / Н.И. Ломакин, М.С. Марамыгин, А.А. Положенцев [и др.] // Международная экономика. — 2023. — № 11. — С. 824–839. — DOI 10.33920/vne-04-2311-06. — EDN NADAIA.

13. L. Tang, Y. Li, G. Yue and D. Li, "Blackhole — A flying paddle algorithm platform and its actual application to quantify the financial sklearn framework", 2021 2nd International Conference on Big Data Economy and Information Management (BDEIM), Sanya, China, 2021, pp. 273–280, doi: 10.1109/BDEIM55082.2021.00062.

14. Маторин, Д.Д. Прогностическое моделирование уровня знаний студентов на основе сочетания мультиномиальной логистической регрессии и нейронных сетей / Д.Д. Маторин, А.Ю. Черепков // Студенческий вестник: актуальные вопросы науки и образования: Сборник студенческих научных работ. — Елец: Елецкий государственный университет имени И.А. Бунина, 2023. — С. 57–62. — EDN TFQJES.

15. G. Ju et al., "A Probabilistic Prediction Method of Carbon Financial Default Risk Based on Python and Machine Learning", 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, United Kingdom, 2023, pp. 369–374, doi: 10.1109/AIARS59518.2023.00081.

16. Gates, Grant, "Microarray Data Analysis and Classification of Cancers" (2019). Williams Honors College, Honors Research Projects. 1025. URL: https://ideaexchange.uakron.edu/honors_research_projects/1025/ (дата обращения: 16.04.2024).

17. S. Maurya and S. Singh, "Time Series Analysis of the Covid-19 Datasets", 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1–6, doi: 10.1109/INOCON50539.2020.9298390.

Zhukova Zhanna Sergeevna
Moscow Technical University of Communications and Informatics, Moscow, Russia
E-mail: assamblee@mail.ru
ORCID: https://orcid.org/0000-0002-9828-4413
RSCI: https://elibrary.ru/author_profile.asp?id=1061900
SCOPUS: https://www.scopus.com/authid/detail.url?authorId=57219179989

Erofeeva Viktoria Vyacheslavovna
Moscow Technical University of Communications and Informatics, Moscow, Russia
Peoples' Friendship University of Patrice Lumumba, Moscow, Russia
E-mail: erofeeva-viktori@mail.ru
ORCID: https://orcid.org/0000-0002-0236-1876

Tagoyev Timur Akhmadjonovich
Moscow Technical University of Communications and Informatics, Moscow, Russia
E-mail: mifers@mail.ru

Ubushayev Nikita Yevgenyevich
Moscow Technical University of Communications and Informatics, Moscow, Russia
E-mail: nikita.ubushaec@gmail.ru

Fyodorov Igor Alexandrovich
Moscow Technical University of Communications and Informatics, Moscow, Russia
E-mail: powerfed@mail.ru

Comparison of temperature forecasting methods based on data from the State of Queensland, Australia

Abstract. The authors present the results of forecasting temperature changes based on an array of data from five meteorological stations in Queensland, Australia. The methods of random forest regression, k-nearest neighbors, linear regression, and seasonal autoregressive integrated moving average were compared to predict temperature changes up to 2030 at different latitudes of the state using regression methods and the seasonal autoregressive integrated moving average time series method. The actual temperature values used for forecasting were collected in two files: «kvin2.xlsx» and «kvshir1.xlsx». «kvin2.xlsx» contains the actual temperatures for five stations for the period from 1856 to 2022, while «kvshir1.xlsx» contains the actual temperatures for 236 stations over the same period. A more detailed comparison of temperature variability from the file «kvin2.xlsx» it was conducted using the random forest method for five stations. In addition, the forecast accuracy for these stations was calculated for two runs, since the prediction used a random spread using the random forest method. Each run of the program outputs new values based on those available in the file «kvin2.xlsx». As a result, predictions were obtained that take into account random variables that are different (but not significantly) for each run of the program. The accuracy was calculated by comparing the predicted temperatures for the two runs with the actual temperatures from the first file. To predict the temperature from a file «kvshir1.xlsx» methods such as k-nearest neighbors, linear regression and seasonal autoregressive integrated moving average were used, without using random scatter. To do this, nine stations with long rows of observations were selected from a variety of stations, since this number is enough to visually demonstrate the work of the program. As a result of comparing different forecasting methods, random forest regressor showed that this method predicts temperature values with an accuracy of at least 96,164 %, the smallest standard error is calculated in the k-nearest neighbors method: 0,175. Based on the random forest regressor, forecasts were made for five stations until 2030.

Keywords: Australia; climate; forecasting; random forest regressor; k-nearest neighbors; linear regression; seasonal autoregressive integrated moving average