

МОДЕЛ ЗА МАШИННО ОБУЧЕНИЕ ЗА КРАТКОСРОЧНА ПРОГНОЗА НА ВРЕМЕТО

Михаил Първанов

*Департамент „Прогнози и информационно обслужване“, Национален институт по метеорология
и хидрология, студент бакалавърска програма „Информатика“ в Нов български университет
e-mail: mparvanovm@gmail.com*

Ключови думи: метеорологичен модел, Random Forest

Резюме: Този доклад представя разработването и прилагането на модел за машинно обучение, базиран на регресия Random Forest, за прогнозиране на краткосрочни тенденции в температурата на въздуха. Моделът е обучен върху многогодишни данни от метеорологични станции, включващи променливи като температура, влажност, скорост на вятъра и атмосферно налягане. Използвани са лагови характеристики и сезонни цикли за улавяне на времевите зависимости. Валидирането показва висока точност при 24-часови прогнози, докато остават предизвикателства при разширяване на прогнозите над 48 часа. Изследването подчертава потенциала на подходите, базирани на изкуствен интелект, да допълват съществуващите числени модели за прогнозиране на времето чрез предоставяне на бързи, основани на данни прогнози с относително ниски изчислителни разходи.

MACHINE LEARNING MODEL FOR SHORT-TERM WEATHER FORECASTING

Mihail Parvanov

*Department of Forecasts and Information Services, National Institute of Meteorology and Hydrology (NIMH),
student in the “Informatics” Bachelor’s program at New Bulgarian University
e-mail: mparvanovm@gmail.com*

Keywords: meteorological model, Random Forest

Abstract: This report presents the development and application of a machine learning model based on Random Forest regression for predicting short-term air temperature trends. The model is trained on multi-year meteorological station data, including variables such as temperature, humidity, wind speed, and pressure. Lagged features and seasonal cycles are incorporated to capture temporal dependencies. Validation demonstrates high accuracy for 24-hour forecasts, while challenges remain for extending predictions beyond 48 hours. The study highlights the potential of AI-based approaches to complement existing numerical weather prediction systems by providing rapid, data-driven insights with relatively low computational costs.

Въведение

Температурата на въздуха е ключов метеорологичен параметър с широко практическо значение. Традиционно прогнозите се изготвят чрез числени модели за прогнозиране на времето (NWP), които решават системи от частни диференциални уравнения, описващи динамиката на атмосферата. Тези подходи са точни, но изискват сериозен изчислителен ресурс и време.

Машинното обучение (ML) предлага алтернативен подход, базиран на извличане на закономерности от исторически наблюдения, без директно решаване на физически уравнения. Настоящото изследване представя регресионен модел Random Forest за краткосрочна прогноза на температура с хоризонт до 72 часа.

Прогнозата на температурата има пряко въздействие върху редица критични сектори. Например, енергийният сектор разчита на краткосрочни прогнози за температурите, за да

оптимизира производството и разпределението на електроенергия и природен газ. В транспорта, температурните колебания влияят върху безопасността на движението, особено през зимния сезон, когато температурите преминават през точката на замръзване. В земеделието температурите определят времето за сеитба, растеж и прибиране на реколтата, както и риска от измръзване или топлинен стрес.

Прогнозите за температурата подпомагат здравните системи при предвиждане на периоди с екстремна горещина или студ. Управлението на рискове при бедствия — например топлинни вълни или студени фронтове — също зависи от надеждни температурни прогнози.

Обсъждане в контекста на съществуващи изследвания

Методът Random Forest има силна теоретична и емпирична подкрепа в научната литература. Още в класическата работа на [1] се подчертава, че ансамбловият подход позволява намаляване на вариацията без увеличаване на пристрастността, което го прави особено подходящ за прогнози в шумни среди като атмосферните процеси.

Класическите статистически методи често се затрудняват с улавянето на нелинейните зависимости между метеорологичните параметри [2]. Random Forest преодолява тази трудност чрез използване на множество дървета и случайност в избора на признаци.

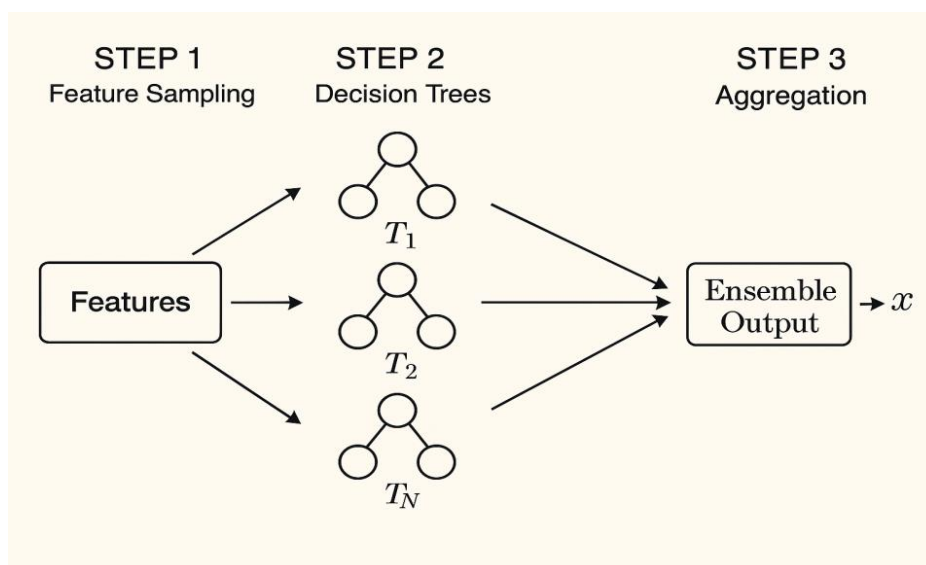
Сравнение с резултатите от числени модели показва, че дори сравнително прости ML модели могат да постигнат сравнима точност в краткосрочен хоризонт. Подобни наблюдения се споменават и в [3], където се подчертава потенциалът на машинното обучение за допълване на NWP моделите.

Предимства и ограничения на Random Forest

Random Forest има няколко ключови предимства:

- Робустност към шум и липсващи данни — моделът запазва добра точност дори при непълни или неточни наблюдения;
- Нелинейно моделиране — не изисква предположения за формата на зависимостта между входа и изхода [4];
- Интерпретируемост — чрез анализ на значимостта на признаците може да се идентифицират ключовите фактори, определящи температурната динамика.
- Въпреки това има и ограничения:
- Невъзможност за екстраполация извън диапазона на наблюденията — моделът се учи от миналото и не може да предвиди стойности в непознати режими;
- Висока зависимост от качеството на входните данни;
- Ограничена физическа интерпретация в сравнение с NWP моделите.

Теоретична основа на метода



Фиг. 1. Схематично представяне на алгоритъма Random Forest

Random Forest е ансамблов модел, който комбинира множество дървета на решение, за да подобри стабилността и точността на прогнозите (фиг. 1). По този начин всяко дърво „вижда“ малко по-различна картина на проблема, което създава разнообразие в ансамбъла. При построяването на всяко дърво на всяка стъпка на разделяне се избира случаен поднабор от входните променливи, от които се търси най-подходящата за разделяне. Това допълнително увеличава разнообразието между дърветата и намалява риска няколко от тях да направят едни и същи грешки.

Параметризацията на сложни атмосферни процеси, като облачна наситеност и турбулентни взаимодействия, е ключов елемент за подобряване на прогностичните модели [5]. Подобни физически механизми могат ефективно да бъдат апроксимирани и от ансамблови машинно-обучаващи се методи като Random Forest.

Всяко дърво извършва прогноза $h_i(x)$ за даден вектор от входни признаци x . Крайната прогноза на ансамбъла се изчислява като средно аритметично (1).

$$(1) \quad \hat{y}(x) = \frac{1}{N} \sum_{i=1}^N h_i(x)$$

След като всички дървета са обучени, техните предсказания се обединяват. При регресионни задачи — резултатите се осредняват, което изглажда индивидуалните грешки на отделните дървета. При класификационни задачи — се използва гласуване по мнозинство. Едно от ключовите предимства на този подход е намаляването на вариацията на модела, без значително увеличаване на отклонението. Индивидуалните дървета могат да бъдат „шумни“ и склонни към преобучаване, но когато се комбинират в ансамбъл, грешките им до голяма степен се компенсират взаимно. Така Random Forest постига висока стабилност и точност, особено при данни с много променливи и нелинейни зависимости. Сложността на модела може да бъде контролирана чрез броя на дърветата в гората и чрез ограничаване на дълбочината им. Колкото повече дървета се използват, толкова по-устойчив става моделът, до момента в който точността се стабилизира. Контролът върху дълбочината на дърветата предотвратява прекомерно напасване върху тренировъчните данни и осигурява по-добра генерализация върху нови наблюдения. Благодарение на тези характеристики, Random Forest се счита за един от най-надеждните ансамблови методи за прогнозиране и често превъзхожда по-опростени модели, особено в реални приложения с шумни и сложни данни.

Обучението цели минимизиране на функцията на загубите, обикновено средноквадратична грешка (2).

$$(2) \quad MSE = \frac{1}{M} \sum_{j=1}^M (y_j - \hat{y}_j)^2$$

Една от силните страни на Random Forest е устойчивостта му спрямо шумни данни и корелации между признаците. Всеки отделен модел — решаващо дърво — е сравнително слаб предсказател, но когато резултатите от стотици такива дървета се усреднят, се получава стабилна и надеждна прогноза. За разлика от линейните модели, Random Forest може да улавя нелинейни зависимости. Важен аспект е и оценката на значимостта на признаците, която помага да се разбере кои фактори влияят най-много върху температурата.

Използвани данни и характеристики

За разработката на модела са осигурени метеорологични данни от отворени локални сензорни станции [6], които предоставят измервания с висока пространствена и времева резолюция (на всеки 3 часа) и обхващат множество наблюдателни точки. Наборът от признаци включва температура, относителна влажност, атмосферно налягане, скорост на вятъра, облачност, валеж, географски координати (ширина, дължина, надморска височина) и синусоидални сезонни признаци (3).

$$(3) \quad \sin\left(\frac{2\pi t}{24}\right), \cos\left(\frac{2\pi t}{24}\right), \cos\left(\frac{2\pi m}{12}\right)$$

Преди обучението на модела данните преминават през няколко етапа на предобработка:

1. Почистване на липсващи стойности чрез интерполация;
2. Нормализация и скалиране до 0–1;
3. Създаване на лагови признаци;
4. Разделяне на тренировъчен и тестов набор по време.

За да се моделира инерцията на температурните процеси, са въведени лагови признаци (4).

$$(4) \quad lag3=T_{(t-3h)}, lag6=T_{(t-6h)}, lag12=T_{(t-12h)}$$

Те са особено ефективни при краткосрочни прогнози и подобряват точността на модела с над 10%.

Обучение и прогнозиране

Моделът е обучен с:

N = 400 решаващи дървета, максимална дълбочина D = 20, минимален брой проби за разделяне min_samples_split = 5. Данните са разделени времево за обучение исторически период до 2023 г.; тест — след 2023 г.

След обучението се извършва прогноза за 72 часа напред (3 дни). За първия ден се използват реални измервания от последните 24 часа, а за втория и третия — усреднени сезонни стойности по град, месец и час.

Метрики за оценка

В оценяването на прогностичните модели ключово значение имат статистическите метрики, които измерват отклонението между прогнозираните и наблюдаваните стойности. Избраните показатели в настоящото изследване — MAE, MSE и RMSE — са сред най-широко използваните в приложната метеорология и численото моделиране [3].

Средна абсолютна грешка (MAE) (5)

$$(5) \quad MAE = \frac{1}{M} \sum_{j=1}^M |y_j - \hat{y}_j|$$

MAE измерва средната абсолютна стойност на разликата между прогнозната температура и реалната наблюдавана температура. Тя предоставя директна и интуитивна интерпретация на грешката, изразена в °C.

При Варна и връх Ботев, MAE е под 2°C за 24–72 часа, което показва изключително добра прогностична точност.

При Пловдив MAE е по-висока, което се дължи на по-резките температурни амплитуди в континенталната част.

Квадратен корен от средноквадратична грешка (RMSE) (6)

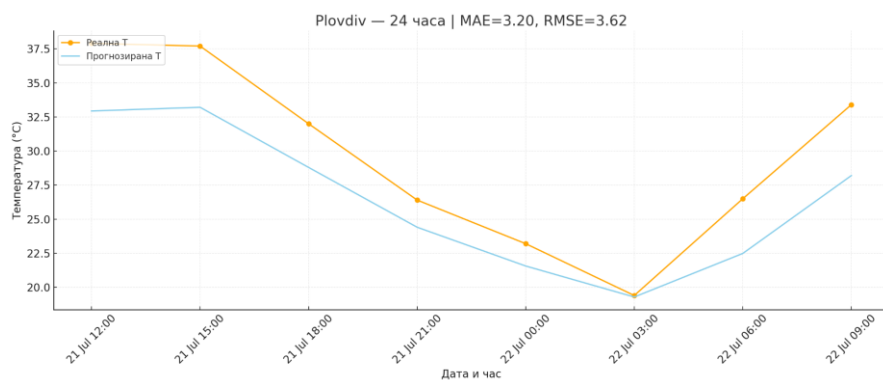
$$(6) \quad RMSE = \sqrt{\frac{1}{M} \sum_{j=1}^M (y_j - \hat{y}_j)^2}$$

RMSE е по-чувствителна към големи отклонения в сравнение с MAE, тъй като изчислява квадратния корен от средната стойност на квадратите на грешките. При анализирани локации RMSE също е най-ниска при морските и планински станции и най-висока във вътрешността на страната.

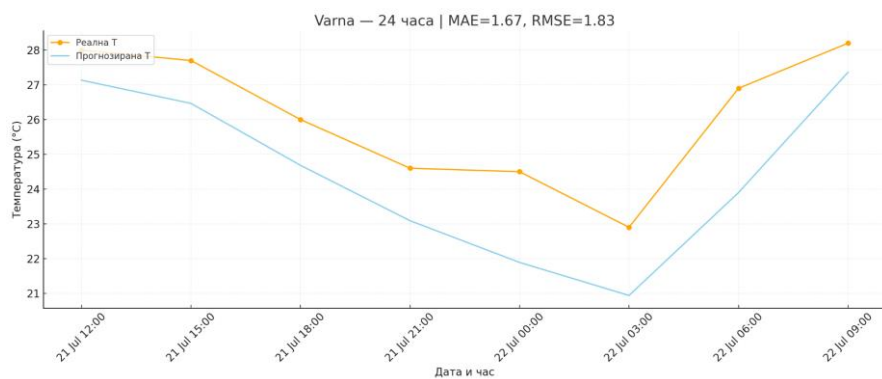
Разликата между RMSE и MAE е малка, което показва липса на сериозни пикови грешки и стабилност на модела. След 48 часа се наблюдава очаквано увеличение на грешките, тъй като с удължаване на хоризонта нараства неопределеността. Грешките при метеорологични процеси нарастват експоненциално с времето поради тяхната хаотична природа.

Четири графика по-долу илюстрират поведението на модела в различни географски и климатични условия:

- Пловдив (24 h) — вътрешен град с по-резки температурни амплитуди (фиг. 2);
- Варна (24 h) — морски климат, плавна температурна крива и ниски грешки (фиг. 3);
- вр. Ботев (72 h) — планински връх, стабилно поведение и отлична точност (фиг. 4);
- София (72 h) — вътрешен град с по-динамичен температурен ход и по-високи грешки (фиг. 5).



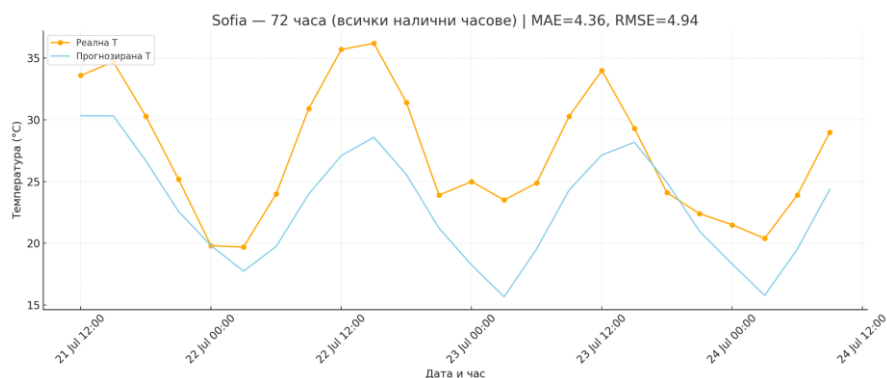
Фиг. 2.



Фиг. 3.



Фиг. 4.



Фиг. 5.

Обобщение на резултатите:

- Точност в рамките на $\pm 2^{\circ}\text{C} \approx 91\%$;
- По-големи отклонения след 48 часа, особено във вътрешността на страната;
- Много добро представяне при морски и планински станции.

Обсъждане на резултатите

Моделът се представя по-добре при крайбрежни станции, където температурните амплитуди са по-ниски и по-плавни. В континенталните райони грешките се увеличават поради по-силно изразени дневни колебания и локални ефекти на релефа. Добавянето на координати и надморска височина подобрява общата стабилност.

Както е показано в предишни числени симулации, промените в температурния профил и концентрацията на CCN оказват съществено влияние върху микрофизиката и динамиката на конвективните облаци [7]. Това частично обяснява по-големите отклонения на прогнозите във вътрешността на страната, където дневните амплитуди са по-изразени.

Числените модели за прогнозиране на времето (NWP) са изключително точни при наличие на суперкомпютърни ресурси. Те обаче изискват огромни обеми данни, сложни физически схеми и значително време за изчисления. Моделът Random Forest, от друга страна, има много по-ниски изчислителни изисквания и дава точни краткосрочни прогнози.

Заклучение

Разработеният модел на базата на Random Forest перспективата показва, че:

- Машинното обучение е ефективен инструмент за краткосрочна прогноза на температура;
- Моделът може да се обучава бързо и да дава резултати с висока точност без нужда от суперкомпютър;
- Подходът е приложим в реално време и може да бъде интегриран като допълнение към съществуващи прогностични системи.

В бъдещи разработки може да се разгледа интеграцията на IWV, извлечен от Глобални навигационни сателитни системи (GNSS), като входна променлива. Подобни подходи са показали добри резултати при анализ на мъгли и нестабилност и могат да осигурят по-богато представяне на влаговия режим в атмосферата, което е критично за температурните прогнози [8].

Комбинирането на статистически модели с данни от числени прогностични системи като ALADIN може да повиши точността на прогнозите, особено при нестабилни атмосферни условия [9]. Това отваря възможности за хибридни подходи между ML и NWP.

Моделът може да бъде внедрен в редица практически системи:

- Енергетика — прогнозиране на пикови натоварвания;
- Здравеопазване — предупреждение за топлинни вълни;
- Земеделие — оптимизация на времето за сеитба;
- Транспорт — оценка на риска от обледяване;
- Климатични услуги — локални прогнози чрез API и мобилни приложения.

Литература:

1. L. Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324
2. D. S. Wilks, Statistical Methods in the Atmospheric Sciences, 3rd ed. Amsterdam, Netherlands: Academic Press, 2011. ISBN: 978-0-12-385022-5
3. S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "WeatherBench: A benchmark dataset for data driven weather forecasting", Journal of Advances in Modeling Earth Systems (JAMES), vol. 12, no. 11, e2020MS002203, 2020. doi: 10.1029/2020MS002203
4. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009. ISBN: 978-0-387-84857-0
5. B. Tsenova, R. Mitzeva, and C. Saunders, "Parametrisation of thunderstorm including the cloud saturation effect", Atmospheric Research, vol. 96, no. 3, pp. 356–365, 2010. doi: 10.1016/j.atmosres.2009.09.012
6. Andonov F., Petrov G. Weather station for smart home applications. Journal – Electrotechnica & Electronica (E+E), Vol. 52 (11-12), 2017, pp. 27–31, ISSN: 0861-4717 (Print), 2603-5421 (Online)
7. B. Tsenova, R. Mitzeva, and A. Brandiyska, "Effect of CCN and temperature profile changes on convective cloud microphysics and dynamics – numerical simulations with 1.5 D cloud model", Bulgarian Journal of Meteorology and Hydrology, submitted, 2012
8. A. Stoycheva and G. Guerova, "Fog case studies using GNSS tropospheric products in Bulgaria", in FIG Working Week 2015: From the Wisdom of the Ages to the Challenges of the Modern World, Sofia, Bulgaria, May 17–21, 2015
9. B. Tsenova and A. Bogatchev, "Evaluation of instability indices computed using ALADIN-BG output and their relation with thunderstorm activity over Bulgaria", in Proceedings of the 16th International Conference on Clouds and Precipitation, Leipzig, Germany, Jul. 2012.