

UDC 504.3.054:004.85

Askarbek Nurislam

Grade 9 student

Supervisor: Yergashbayev Zhasur

computer science teacher

Astana Bilim-Innovation Lyceum for Boys

(Astana, Kazakhstan)

CAN THE WEATHER TELL WHICH DAYS IN ASTANA WILL BE POLLUTED? CLASSIFYING PM_{2.5} EXCEEDANCE DAYS WITH MACHINE LEARNING AND CHECKING THE DATA AGAINST A GROUND MONITOR

Abstract: Astana has long cold winters, and part of the city is still heated with coal. In this study I asked whether weather data alone can show in advance on which days the daily mean concentration of fine particles (PM_{2.5}) in Astana will be above the World Health Organization guideline of 15 $\mu\text{g}/\text{m}^3$. I used free hourly PM_{2.5} estimates from the Copernicus Atmosphere Monitoring Service (CAMS) and ERA5 weather data for 1,490 days between 4 August 2022 and 1 September 2026. Only 141 days (9.5%) were above the guideline, so a model that always answers "normal" is right 90.5% of the time and still finds no polluted day. I trained logistic regression, random forest and gradient boosting models on 2022-2024 and tested them once on 2025-2026. With the weather of the same day, the random forest found 64% of the exceedance days with 55% of its alarms correct ($F1 = 0.59$, 95% CI 0.47-0.68), better than the rule "tomorrow will be like today" ($F1 = 0.42$). When the model could use only the weather of the previous day, $F1$ fell to 0.32, below that simple rule. The model missed all ten summer exceedances and most days with a lot of dust in the air. Finally, I compared CAMS with the US Embassy monitor in Astana on 613 days. The monitor showed about twice as much PM_{2.5} (17.4 against 8.4 $\mu\text{g}/\text{m}^3$ on average) and exceedances on 39% of days instead of 9%, and only the monitor showed higher pollution in the heating season. Weather predicted the monitor's exceedance days at least as well as the CAMS ones. My conclusion is that weather contains real information about polluted days in Astana, but CAMS shows a much cleaner city than the ground monitor, so models trained only on CAMS should be used with care.

Keywords: PM_{2.5}; air quality; Astana; machine learning; imbalanced classification; CAMS; ground-based monitoring

1. Introduction

I live in Astana, and every winter there are evenings when the air smells of smoke and the city looks grey. Part of the city is still heated with coal, and the heating season lasts about half a year. Air pollution in Astana and its monitoring system have been studied before [1]. Studies describe coal heating and thermal power plants close to the city as important sources of particulate pollution [2], and a recent sampling study found soot and coal fly ash in Astana's particulate matter and PM_{2.5} concentrations up to 534 $\mu\text{g}/\text{m}^3$ in individual samples [3].

PM_{2.5} means particles with a diameter of 2.5 micrometres or less. They go deep into the lungs, and in 2021 the World Health Organization lowered its guideline for the 24-hour mean of PM_{2.5} to 15 $\mu\text{g}/\text{m}^3$ [4]. For a person living in the city, the exact number is less useful than a simple answer to the question "will tomorrow be a bad day?". This is why I decided to treat the

problem as a classification problem with two classes, "normal" and "exceedance", instead of predicting the exact concentration.

Predicting PM_{2.5} with machine learning is a common topic. Studies from Iran used meteorological data and several algorithms to predict PM_{2.5} levels or concentrations [5, 6], and a study in *Frontiers in Big Data* classified PM_{2.5} into several groups and used oversampling because the classes were imbalanced [7]. Imbalance is a known problem in air quality classification [8]. When one class is rare, accuracy is misleading, and precision-recall curves and the Matthews correlation coefficient describe a model better [9, 10].

Most such studies use measurements from ground stations. For Astana I could not find a long and complete public record of measurements, so my plan was to use the free PM_{2.5} estimates of the Copernicus Atmosphere Monitoring Service (CAMS) through the Open-Meteo service [11, 12]. CAMS values are not measurements. They come from an atmospheric model that combines emissions, meteorology and satellite observations [13]. Studies that compared CAMS with ground stations over China [14], around the world using the OpenAQ platform [15] and in Hungary [16] show that CAMS PM_{2.5} should be checked before use, and in Hungary the raw CAMS data systematically underestimated the measured values [16]. There is also a risk of circular reasoning. CAMS itself uses weather, so a model that predicts CAMS PM_{2.5} from weather may only learn how CAMS works, not how pollution in the real city behaves.

I did not find a public study for Astana that treats the problem as classification of exceedance days, compares the model with honest baselines on imbalanced data, and checks the data source against a ground monitor. My search was in English and Russian, and I do not claim it was complete. In this study I asked four questions:

- How much better than simple baselines can weather-based models find the days when PM_{2.5} in Astana is above 15 µg/m³?
- Which weather features are the most important, and does this agree with what is known about how pollution accumulates and disperses?
- What do the days on which the model is wrong have in common?
- How well do the CAMS values agree with a ground monitor in Astana?

Before the analysis I wrote four hypotheses. H1: a model with only weather features will beat the baseline "always normal" in F1 for the exceedance class, but not necessarily in accuracy. H2: the most important features will be those connected to dispersion, such as low wind speed, wind direction and stagnant conditions. H3: missed exceedances will mostly be days when the pollution was caused by something other than the weather, which CAMS smooths out. H4: CAMS will agree with the ground monitor only weakly, especially at low concentrations.

2. Data and methods

PM_{2.5}. I downloaded hourly PM_{2.5} for the point 51.18° N, 71.45° E from the Open-Meteo Air Quality API [12]. For Astana this API gives the CAMS global atmospheric composition forecast with a resolution of 0.4° (about 45 km), which is available from August 2022. This explains why my data start on 4 August 2022. At first I thought the data were a reanalysis, but the documentation shows that for Central Asia they are archived forecasts, so I call them "CAMS estimates". All hours are in the fixed time zone GMT+5, and each day's mean is the average of 24 hourly values. The first day, 4 August 2022, has 19 hours and I kept it, so the full set has 1,490 days up to 1 September 2026.

Weather. Hourly weather came from the ERA5 reanalysis [17] through the Open-Meteo Historical Weather API [11], for the nearest ERA5 grid point (51.25° N, 71.5° E). I used air

temperature, relative humidity, surface pressure, wind speed and direction at 10 m, precipitation and cloud cover. I requested only ERA5, so that all years come from one model. Open-Meteo also gives the boundary layer height, which is physically very important for pollution, but in this archive it had no values from January to June 2024, so I did not use it in the models.

Ground monitor. The US Embassy in Astana had a PM_{2.5} monitor whose data were published through the AirNow programme. I could not find this monitor in the OpenAQ search. Its records are in the open OpenAQ data archive as location 7094, "US Diplomatic Post: Astana", and they can be downloaded without registration [18]. Hours with the code -999 were treated as missing. The time stamps in the files change from UTC+6 to UTC+5 in March 2024, when Kazakhstan moved to a single time zone, so I converted every hour to UTC first and then to GMT+5. A station day was used only if it had at least 18 valid hours.

Target and features. A day was an exceedance ($y = 1$) if its daily mean CAMS PM_{2.5} was above $15 \mu\text{g}/\text{m}^3$. From the hourly weather I calculated daily mean, minimum and maximum temperature and their range, mean humidity, mean pressure and its change from the previous day, mean and maximum wind speed, the number of calm hours (wind below 2 m/s), wind direction as the sine and cosine of the vector-mean direction, total precipitation and mean cloud cover. I also added the previous day's temperature and wind speed, the change in temperature, the month (as sine and cosine) and a heating season flag (October-March).

Three modes. In mode A the model gets the weather of the same day. This is the setting of my research question, and it is also the most optimistic one, because real forecasters know tomorrow's weather only from a forecast. In mode B the model also gets the previous day's PM_{2.5}, which makes the task easier because pollution changes slowly. I added mode P after the first results: the model gets only the weather of the previous day and the calendar, which is what is really known on the evening before.

Splitting the data. Neighbouring days are similar, so a random split would let the model see the test days through their neighbours. I used a chronological split: training on 4 August 2022 - 31 December 2024 (880 days after removing the first day, which has no previous day, 67 exceedances) and testing on 1 January 2025 - 1 September 2026 (609 days, 74 exceedances). Hyperparameters were chosen by expanding-window cross-validation inside the training part, using the area under the precision-recall curve (PR-AUC). The decision threshold was also chosen inside the training part, as the threshold that gave the best F1 on the cross-validation predictions. The test part was used only once.

Models. I used two baselines: "always normal" and the persistence rule "today is an exceedance if yesterday was". The models were logistic regression with class weights (with standardised features), random forest [19] with balanced class weights, and gradient boosting with balanced sample weights. Everything was done in Python with pandas and scikit-learn [20], with a fixed random seed.

Metrics. For the exceedance class I report precision, recall, F1, PR-AUC and the Matthews correlation coefficient (MCC) [10]. PR-AUC is more informative than ROC-AUC when one class is rare [9]. The test part has only 74 exceedance days, so I calculated 95% confidence intervals with a block bootstrap: 2,000 resamples made of blocks of 7 consecutive days, so that the dependence between neighbouring days is kept. To compare two models, I calculated the difference in F1 on the same resamples.

Error analysis. For the main model I divided the test days into found exceedances (TP), missed exceedances (FN), false alarms (FP) and correct normal days (TN), and compared their

weather, season, position in an episode of polluted days and the amount of dust in the CAMS data. I compared TP and FN with the Mann-Whitney test.

3. Results

3.1 What the CAMS data show

The daily mean CAMS PM_{2.5} for the whole period was 8.73 $\mu\text{g}/\text{m}^3$ (median 7.83, standard deviation 5.27). The 90th percentile was 14.8 $\mu\text{g}/\text{m}^3$, almost exactly at the threshold, which means that exceedances are rare by definition. The highest daily mean was 49.9 $\mu\text{g}/\text{m}^3$ on 28 November 2024, and the highest hourly value was 65.9 $\mu\text{g}/\text{m}^3$ on the same day (Figure 1). Only 141 of 1,490 days (9.5%) were above 15 $\mu\text{g}/\text{m}^3$, so the "always normal" model has an accuracy of 90.5% on the full data and 87.8% on the test part.

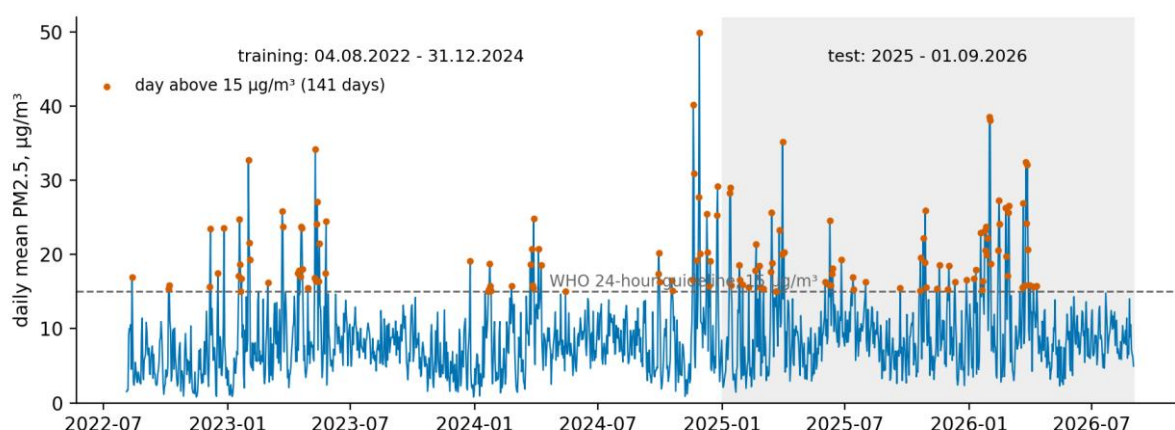


Figure 1. Daily mean CAMS PM_{2.5} in Astana, 4 August 2022 - 1 September 2026. Orange points are days above the WHO guideline; the grey area is the test period

When I first counted exceedances by calendar year, it looked like the share of exceedance days grew more than twice, from 4.7% in 2022 to 11.8% in 2025 (Table 1). This comparison was wrong, because 2022 starts in August and misses January-March, which in these data are some of the most polluted months. When I compared only August-December, which all years have, 2023 turned out to be the cleanest year (0.7%), then there was a jump in 2024 (11.8%) and a small fall in 2025 (9.2%). So the honest description is a jump between 2023 and 2024 without a clear trend after it.

Table 1. CAMS PM_{2.5} exceedance days in Astana by year

Year	Days	Exceedance days	Share, %	Mean PM _{2.5} , $\mu\text{g}/\text{m}^3$	Aug-Dec share, %
2022	150	7	4.7	5.95	4.7
2023	365	28	7.7	8.24	0.7
2024	366	32	8.7	8.80	11.8
2025	365	43	11.8	9.28	9.2
2026	244	31	12.7	10.22	-

2022 starts on 4 August and 2026 ends on 1 September, so the yearly shares cannot be compared directly; the last column uses only August-December.

In the CAMS data the heating season was almost as clean as the warm season (8.81 against 8.64 $\mu\text{g}/\text{m}^3$), and the most polluted month was March, which also had the most dust in the CAMS data (12.4 $\mu\text{g}/\text{m}^3$ on average, compared with about 1 $\mu\text{g}/\text{m}^3$ in August and

December). I expected the opposite, and section 3.5 shows that the ground monitor disagrees with CAMS here.

3.2 How well the models find exceedance days

Table 2 shows the results on the test period. The "always normal" model has an F1 of zero. The persistence rule found 42% of the exceedance days, with 42% of its alarms correct (F1 = 0.42).

Table 2. Results on the test period (609 days, 74 exceedance days)

Mode and model	Precision	Recall	F1 [95% CI]	PR-AUC [95% CI]	MC C
Always "normal"	-	0.00	0.00	-	0.00
Same as yesterday	0.42	0.42	0.42 [0.30-0.51]	0.46 [0.33-0.58]	0.34
A: logistic regression	0.30	0.82	0.44 [0.34-0.53]	0.51 [0.36-0.66]	0.39
A: random forest	0.55	0.64	0.59 [0.47-0.68]	0.54 [0.39-0.70]	0.53
A: gradient boosting	0.41	0.53	0.46 [0.34-0.56]	0.45 [0.32-0.60]	0.38
B: logistic regression	0.48	0.59	0.53 [0.42-0.62]	0.60 [0.47-0.71]	0.46
B: random forest	0.61	0.61	0.61 [0.49-0.70]	0.69 [0.56-0.79]	0.55
B: gradient boosting	0.53	0.66	0.59 [0.48-0.67]	0.58 [0.43-0.70]	0.53
P: logistic regression	0.23	0.69	0.34 [0.24-0.43]	0.30 [0.20-0.43]	0.24
P: random forest	0.21	0.66	0.32 [0.23-0.41]	0.32 [0.21-0.45]	0.22
P: gradient boosting	0.18	0.78	0.30 [0.22-0.38]	0.23 [0.15-0.33]	0.20

A: weather of the same day; B: A plus yesterday's PM2.5; P: only yesterday's weather and the calendar. Confidence intervals are from a 7-day block bootstrap. For the persistence rule PR-AUC uses yesterday's PM2.5 as the score.

In mode A the random forest was the best model. It found 47 of the 74 exceedance days, gave 39 false alarms and missed 27 exceedances, and its accuracy was 89.2%. This is a little higher than 87.8% for "always normal", but the important difference is in F1: 0.59 against 0. The random forest was also better than the persistence rule (difference in F1 0.17, 95% CI 0.05-0.28) and better than logistic regression (difference 0.15, CI 0.06-0.23). Logistic regression found the most exceedances (82%), but it gave 141 false alarms, so only 30% of its alarms were right. Gradient boosting was not better than logistic regression.

Adding yesterday's PM2.5 (mode B) raised PR-AUC from 0.54 to 0.69 for the random forest, but the F1 improvement was small and not clear (difference 0.02, CI -0.07 to 0.10). For logistic regression the gain from mode B was clearer (difference 0.09, CI 0.003-0.17).

The result that surprised me most was mode P. When the models could use only the weather of the day before, F1 fell to 0.30-0.34, and the random forest was worse than the simple persistence rule (difference -0.10, CI -0.18 to -0.005). Figure 2 shows the same thing with

precision-recall curves: the curve for mode P lies far below the curves for modes A and B. So most of the skill of mode A comes from knowing the weather of the same day.

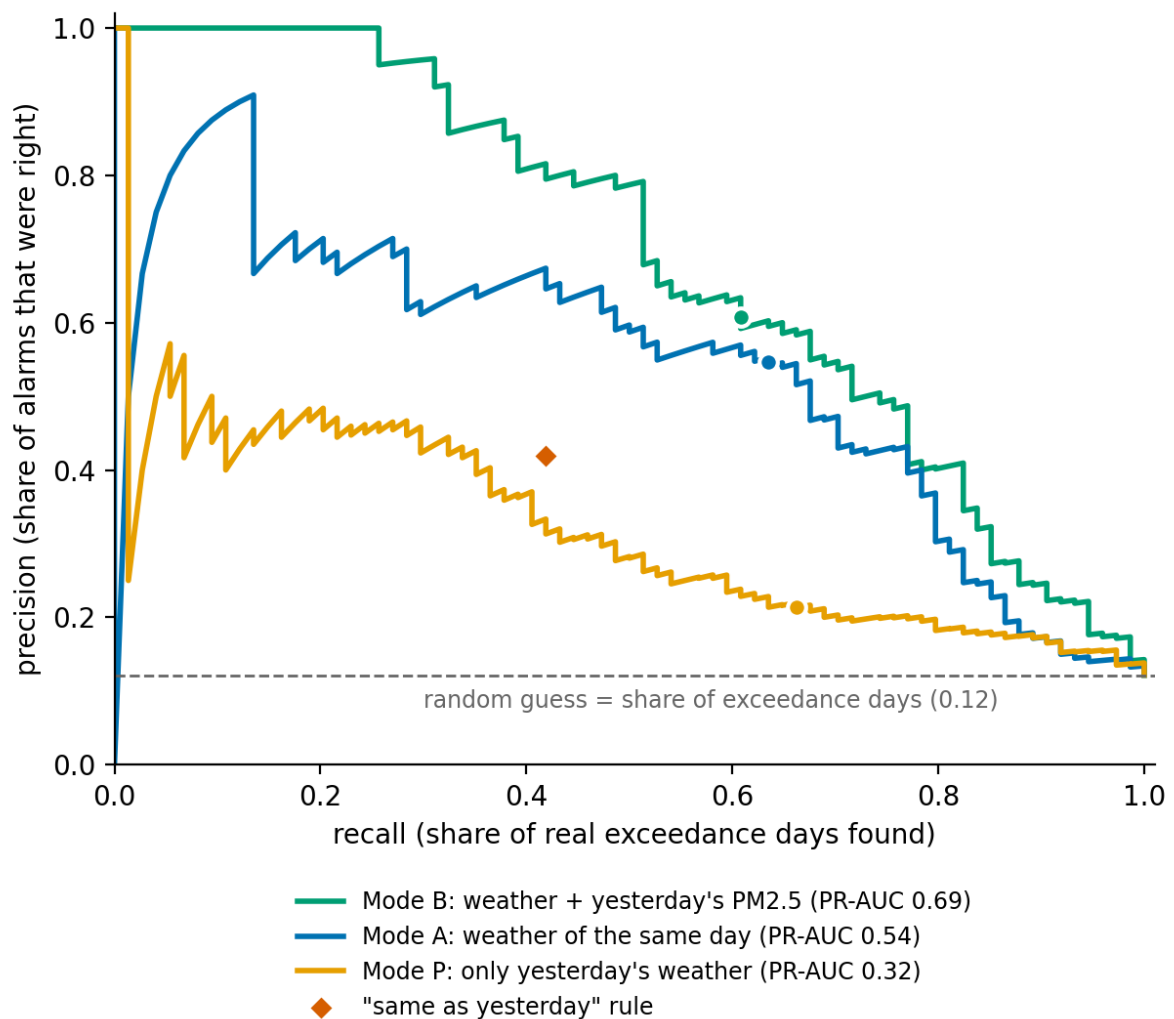


Figure 2. Precision-recall curves of the three random forest models on the test period. Points show the chosen thresholds; the diamond is the "same as yesterday" rule

One more thing I noticed is that cross-validation inside the training part gave much lower PR-AUC values (0.14-0.19 in mode A) than the test. The training years had fewer exceedances, especially 2023 with only one in August-December, so some validation folds had very few positive days. The test result may therefore be a bit lucky, and I do not think the exact numbers would repeat in another year.

3.3 Which features matter

In the random forest (mode A), the features that most reduced PR-AUC when they were shuffled were the maximum wind speed of the day, the previous day's temperature, the minimum temperature and the mean wind speed. Wind direction had only a small effect. In logistic regression, the month, more calm hours, higher pressure and a larger daily temperature range increased the chance of an exceedance, and stronger wind decreased it. Many of these features are correlated with each other, so I read the signs of the coefficients only as a rough picture. In mode B yesterday's PM2.5 was the most important feature, and after it again wind and

temperature. This agrees with the idea that pollution builds up when the air is calm and cold, and it partly supports H2, except that wind direction was not important.

3.4 Where the model makes mistakes

The 27 missed exceedances were different from the 47 found ones. On the missed days the median wind speed was higher (3.4 against 2.3 m/s, $p < 0.001$), there were no calm hours (median 0 against 7), the median temperature was 5.1 °C instead of -11.4 °C, and the ratio of dust to PM10 in the CAMS data was about twice as high (0.21 against 0.10, $p = 0.003$). In other words, the model learned that a polluted day is a cold and calm day, and it missed the polluted days that did not look like that. The position in an episode did not explain the misses: 56% of the missed days and 60% of the found days were the first day of a polluted period.

Three patterns explain most misses. First, many misses were close to the threshold: 63% of the missed days had PM2.5 between 15 and 18 $\mu\text{g}/\text{m}^3$, compared with 34% of the found days, so a part of the errors comes from the sharp threshold itself. Second, the model did not find any of the 10 exceedances in June-August 2025. The training years had only one summer exceedance, so the model had nothing to learn from. On several of these June days the CAMS aerosol optical depth was 0.4-0.5, about twice its usual summer value of 0.19, which points to aerosol brought from somewhere else, but from my data I cannot say whether it was dust or smoke. Third, on days when CAMS dust was above 10 $\mu\text{g}/\text{m}^3$, the model found only 7 of 18 exceedances, while on the other days it found 40 of 56.

The 39 false alarms were mostly cold, calm days in winter and spring when PM2.5 stayed below the threshold. Their median PM2.5 was 10.7 $\mu\text{g}/\text{m}^3$, and 36% of them were above 12 $\mu\text{g}/\text{m}^3$, so many false alarms were "almost exceedances".

This means that H3 was only partly right. The missed days were not explained by local sources, because CAMS does not see local sources anyway. They were days when the particles probably came from outside the city, or days very close to the threshold.

3.5 CAMS against the ground monitor

The US Embassy monitor had valid data from 20 October 2022 to 4 March 2025, but with gaps. There were 613 days (24 October 2022 - 3 March 2025) with at least 18 hours of measurements, which is 65% of the CAMS days in this period. Table 3 compares the two sources on these days.

Table 3. CAMS and the US Embassy monitor on the same 613 days

Measure	CAMS	US Embassy monitor
Mean PM2.5, $\mu\text{g}/\text{m}^3$	8.4	17.4
Median PM2.5, $\mu\text{g}/\text{m}^3$	7.5	12.6
Highest daily mean, $\mu\text{g}/\text{m}^3$	49.9	117.8
Share of exceedance days, %	8.8	39.0
Heating season mean (Oct-Mar), $\mu\text{g}/\text{m}^3$	7.8	20.4
Warm season mean (Apr-Sep), $\mu\text{g}/\text{m}^3$	9.1	13.8

The highest CAMS value is for the whole period; the monitor's highest value was on 6 December 2022. Days are in GMT+5.

The two sources moved together only moderately (Pearson $r = 0.49$, Spearman $\rho = 0.43$). On days when the monitor was below 15 $\mu\text{g}/\text{m}^3$ the correlation was only 0.28, which agrees with H4. CAMS was lower on average by 9.0 $\mu\text{g}/\text{m}^3$, and the difference was largest in the heating season (Figure 3a). The monitor recorded an exceedance on 239 days, and CAMS also

showed an exceedance on only 50 of them. CAMS showed an exceedance without the monitor on 4 days. The two sources agreed on 68.5% of days, and Cohen's κ was 0.23.

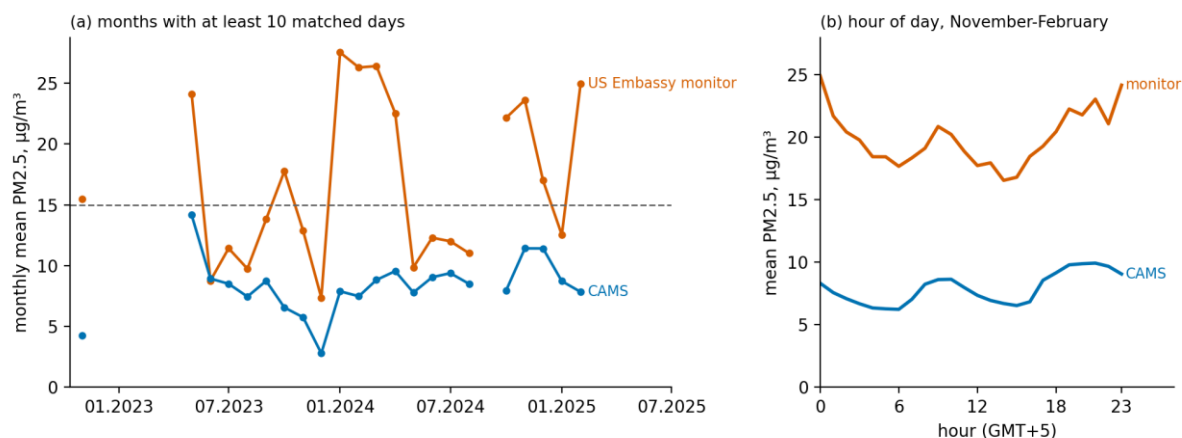


Figure 3. CAMS and the US Embassy monitor: (a) monthly means on matched days; (b) mean hour-of-day profile in November-February

The most important difference is the season. In the CAMS data the heating season was cleaner than the warm season, but at the monitor it was about one and a half times more polluted (20.4 against 13.8 $\mu\text{g}/\text{m}^3$), and half of the heating season days were above the guideline. In winter, CAMS shows two peaks, around 9-10 in the morning and 19-21 in the evening. The monitor also has a small morning peak, but its highest values are around midnight, and the night stays polluted (Figure 3b). This fits the picture of houses burning coal in the evening while cold, stable air holds the smoke near the ground.

3.6. Does weather explain the real monitor as well?

To check the circularity problem, I trained the same random forest (mode A) on the 613 matched days twice, once with the CAMS exceedances as the target and once with the monitor's exceedances, using the same expanding-window cross-validation. Because the share of positive days was very different (9% and 39%), I compared ROC-AUC, which does not depend on this share. It was 0.78 for CAMS and 0.82 for the monitor. So the weather predicted the real measurements at least as well as the CAMS estimates.

The relationships were not the same, though. Wind speed and calm hours were related to PM2.5 in both sources in a similar way. Temperature was not: the monitor's PM2.5 was higher on colder days (Spearman $\rho = -0.30$), while CAMS PM2.5 was slightly higher on warmer days ($\rho = 0.15$). Pressure ($\rho = 0.44$ against 0.06) and precipitation ($\rho = -0.42$ against -0.18) were also much more strongly related to the monitor than to CAMS.

4. Discussion

H1 was supported in mode A: the random forest clearly beat both baselines in F1, while its accuracy was only a little higher than that of "always normal". Mode P shows the limit of this result. With only yesterday's weather, the models were worse than the persistence rule, so in a real forecast the quality would depend mostly on the quality of the weather forecast for the next day. My mode A is closer to the case of a perfect weather forecast, and the real result would be somewhere between modes A and P.

H2 was partly supported. Wind speed, calm hours and low temperature were the most important features, but wind direction was not. Maybe the 45 km CAMS cell is too large for the direction of local sources to matter.

H3 was not supported in the form I wrote it. I expected missed days to be caused by local sources that CAMS smooths out, but local sources are missing from the CAMS target as well, so the model could not miss them. The missed days were warm, windy and often dusty days, or days just above $15 \mu\text{g}/\text{m}^3$. This shows that a "weather only" model does not know where particles come from, and that one threshold creates many borderline errors.

H4 was supported, and the comparison with the monitor was the most important part of this work for me. CAMS showed about half of the PM_{2.5} measured at the US Embassy, and it showed four times fewer exceedance days. The finding that the heating season was not more polluted, which I wrote in my first draft, turned out to come from the CAMS data and not from the real city. A model trained on CAMS learns the rule "exceedances happen on calm days, and dust can add some", while the monitor shows that in Astana cold winter days are polluted much more often. At the same time, the circularity worry was not confirmed: weather explained the monitor's exceedances at least as well as the CAMS ones. I think the right conclusion is that weather data are useful for this task, but the target should be measured PM_{2.5}, and CAMS should be calibrated against stations before it is used for Astana, as was done for Hungary [16].

During the work I also learned two practical lessons. The first is that comparing incomplete years with full years can create a false trend, as in Table 1. The second is that time zones matter. Kazakhstan changed its time zone in 2024, and if I had used the station's local hours without converting them, the hour-of-day profiles would be shifted by one hour for part of the period.

Limitations. The main limitation is that the model's target is a CAMS estimate, which section 3.5 shows to be far below the measured values. The test period has only 74 exceedance days, so the confidence intervals are wide, and the cross-validation results inside the training part were much weaker than the test results. The data cover only four years, and the jump between 2023 and 2024 shows that conditions change from year to year. The comparison uses one monitor at one place in the city, which had data on 65% of the days and stopped publishing on 4 March 2025, so it does not cover the test period. I used observed weather, not weather forecasts. I did not use emissions, traffic or the boundary layer height, and I tested only one threshold ($15 \mu\text{g}/\text{m}^3$).

5. Conclusion

Using 1,490 days of CAMS and weather data for Astana, I showed that a random forest using the weather of the same day can find about two thirds of the days with PM_{2.5} above the WHO guideline, with about half of its alarms correct, which is clearly better than "always normal" and the persistence rule. With only the previous day's weather, the models lost this advantage. The model missed summer exceedances and dusty days, which were rare in the training years. The comparison with the US Embassy monitor showed that CAMS greatly underestimates PM_{2.5} in Astana, especially in the heating season, and finds only about one in five of the exceedance days that the monitor recorded. For the next step I would like to train the same models on measured PM_{2.5} from Astana's ground stations, include the boundary layer height and real weather forecasts, and check whether a model trained on measured data can warn people about polluted days one day in advance.

Author contribution and use of AI

The author chose the topic, formulated the research question and hypotheses, wrote the research plan, downloaded the CAMS data, calculated the yearly statistics and the August-December comparison, found the problem with comparing incomplete years, described the hour-of-day pattern and the circularity problem, and collected the first list of literature. The

author's mentor found the US Embassy monitor in the OpenAQ archive, built the models and the comparison with the monitor, made the figures, and helped write and edit the English text. Generative AI tools were used to help with editing and translation. Every number in the paper was calculated by the analysis code from the open data sources and checked by the author's mentor.

References:

1. Kerimray A., Bakdolotov A., Sarbassov Y., Inglezakis V., Pouloupoulos S. Air pollution in Astana: analysis of recent trends and air quality monitoring system // *Materials Today: Proceedings*. 2018. Vol. 5, No. 11. P. 22749-22758. DOI: 10.1016/j.matpr.2018.07.086.
2. Biloshchytskyi A., Neftissov A., Kuchanskyi O., Andrashko Y., Biloshchytska S., Mukhatayev A., Kazambayev I. Fractal analysis of air pollution time series in urban areas in Astana, Republic of Kazakhstan // *Urban Science*. 2024. Vol. 8, No. 3. Art. 131. DOI: 10.3390/urbansci8030131.
3. Agibayeva A., Kumisbek A., Nauyryzbay A., Avcu E., et al. Towards sustainable air quality in coal-heated cities: a case study from Astana, Kazakhstan // *Sustainability*. 2025. Vol. 17, No. 22. Art. 10214. DOI: 10.3390/su172210214.
4. World Health Organization. WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization, 2021. 273 p. ISBN 978-92-4-003422-8. URL: <https://www.who.int/publications/i/item/9789240034228> (accessed: 02.10.2026).
5. Mohammadi F., Teiri H., Hajizadeh Y., Abdolahnejad A., Ebrahimi A. Prediction of atmospheric PM_{2.5} level by machine learning techniques in Isfahan, Iran // *Scientific Reports*. 2024. Vol. 14. Art. 2109. DOI: 10.1038/s41598-024-52617-z.
6. Makhdoomi A., Sarkhosh M., Ziaei S. PM_{2.5} concentration prediction using machine learning algorithms: an approach to virtual monitoring stations // *Scientific Reports*. 2025. Vol. 15. Art. 8076. DOI: 10.1038/s41598-025-92019-3.
7. Saminathan S., Malathy C. Ensemble-based classification approach for PM_{2.5} concentration forecasting using meteorological data // *Frontiers in Big Data*. 2023. Vol. 6. Art. 1175259. DOI: 10.3389/fdata.2023.1175259.
8. Ravindiran G., Karthick K., Sivarethinamohan S., Rajamanickam S., et al. A cluster-aware synthetic resampling and machine learning framework for multi-class Air Quality Index classification // *Environmental Monitoring and Assessment*. 2026. Vol. 198, No. 3. Art. 270. DOI: 10.1007/s10661-026-15121-7.
9. Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets // *PLOS ONE*. 2015. Vol. 10, No. 3. Art. e0118432. DOI: 10.1371/journal.pone.0118432.
10. Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation // *BMC Genomics*. 2020. Vol. 21. Art. 6. DOI: 10.1186/s12864-019-6413-7.
11. Zippenfenig P. Open-Meteo.com Weather API [software]. Zenodo, 2024. DOI: 10.5281/zenodo.7970649.

12. Open-Meteo. Air Quality API: CAMS global atmospheric composition forecasts, 0.4°, from August 2022. URL: <https://open-meteo.com/en/docs/air-quality-api> (accessed: 02.10.2026).
13. Rémy S., Kipling Z., Huijnen V., Flemming J., et al. Description and evaluation of the tropospheric aerosol scheme in the Integrated Forecasting System (IFS-AER, cycle 47R1) of ECMWF // Geoscientific Model Development. 2022. Vol. 15, No. 12. P. 4881-4912. DOI: 10.5194/gmd-15-4881-2022.
14. Ali M.A., Bilal M., Wang Y., Nichol J.E., et al. Accuracy assessment of CAMS and MERRA-2 reanalysis PM_{2.5} and PM₁₀ concentrations over China // Atmospheric Environment. 2022. Vol. 288. Art. 119297. DOI: 10.1016/j.atmosenv.2022.119297.
15. Jin C., Wang Y., Li T., Yuan Q. Global validation and hybrid calibration of CAMS and MERRA-2 PM_{2.5} reanalysis products based on OpenAQ platform // Atmospheric Environment. 2022. Vol. 274. Art. 118972. DOI: 10.1016/j.atmosenv.2022.118972.
16. Qor-el-aïne A., Béres A., Géczi G. Calibration of CAMS PM_{2.5} data over Hungary: a machine learning approach // Environmental Research Communications. 2024. Vol. 6, No. 7. Art. 075026. DOI: 10.1088/2515-7620/ad6239.
17. Hersbach H., Bell B., Berrisford P., Hirahara S., et al. The ERA5 global reanalysis // Quarterly Journal of the Royal Meteorological Society. 2020. Vol. 146, No. 730. P. 1999-2049. DOI: 10.1002/qj.3803.
18. OpenAQ. Open data archive, location 7094 "US Diplomatic Post: Astana" (AirNow), hourly PM_{2.5}, 2022-2025. URL: <https://openaq-data-archive.s3.amazonaws.com> (accessed: 02.10.2026).
19. Breiman L. Random forests // Machine Learning. 2001. Vol. 45, No. 1. P. 5-32. DOI: 10.1023/A:1010933404324.
20. Pedregosa F., Varoquaux G., Gramfort A., Michel V., et al. Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825-2830.

МАЗМҰНЫ – СОДЕРЖАНИЕ – CONTENT

ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ - СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ - SOCIO-ECONOMIC AND LEGAL SCIENCES

Төлегенов Тимур Сейфуллаулы ДИНАМИКА ТУРИСТСКИХ ПОТОКОВ И ЭФФЕКТИВНОСТЬ ИНФРАСТРУКТУРЫ ТУРКЕСТАНА: ОТ ЭКСТЕНСИВНОГО РОСТА К СМАРТ-УПРАВЛЕНИЮ	10
Alexander Gorevoy WHICH INEQUALITY STATISTICS SURVIVE AN ADMINISTRATIVE BOUNDARY CHANGE? EVIDENCE FROM KAZAKHSTAN, 2010–2023	14
Rakhimov Djalil DOES THE SHIFT TO CASHLESS PAYMENTS GO TOGETHER WITH FASTER GROWTH IN HOUSEHOLD CREDIT? EVIDENCE FROM THE REGIONS OF KAZAKHSTAN, 2023-2025	26
Kazhgerey Asan HOW MANY FOOD BASKETS DOES AN AVERAGE WAGE BUY? WAGES AND FOOD PRICES IN THE 20 REGIONS OF KAZAKHSTAN, 2024-2026	37
Tagimova Alua WHO WAS LEFT BEHIND? EDUCATION AND FINANCIAL INCLUSION IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024	44
Wei Wei, Mynzhanova Gulzhakan, Niyetalina Gaukhar DIGITAL–PHYSICAL INTEGRATION AND ECONOMIC PERFORMANCE IN INDUSTRIAL INTEGRATED ENERGY SYSTEMS: A SCENARIO-BASED ANALYSIS OF SMART GRID CONTROL, BATTERY ENERGY STORAGE, AND DIGITAL TWINS	51
Xu Fumin, Wei Wei, Ruziyeva Elvira, Moldabekova A. FROM CULTURAL RESOURCES TO GLOBAL CONSUMER VALUE: THE BRIDGE–IMC MODEL AND THE INTERNATIONALIZATION OF BLACK MYTH: WUKONG UNDER THE BELT AND ROAD INITIATIVE	64
Ibraimkhan Alikhan WHICH CRIMES CAN A 14-YEAR-OLD BE TRIED FOR? THE LIST IN ARTICLE 15 OF KAZAKHSTAN'S CRIMINAL CODE COMPARED WITH 30 OTHER JURISDICTIONS	80
Yerboluly Madi DID IMPORTED CARS REALLY GET CHEAPER? CHINESE CARS AND THE AVERAGE CUSTOMS VALUE OF PASSENGER CARS IMPORTED INTO KAZAKHSTAN, 2015-2025	91
Салимбаев Ербол Аманжолович АНАЛИЗ ИНВЕСТИЦИОННОЙ АКТИВНОСТИ КАК ИНДИКАТОР ИЗМЕНЕНИЙ ПОТЕНЦИАЛЬНОГО СПРОСА НА КОНСАЛТИНГОВЫЕ УСЛУГИ	99
Rambergenov Ramil REGIONAL NEW-FIRM FORMATION IN UZBEKISTAN AFTER THE 2017 REFORMS: PERSISTENCE IN SCALE, CONVERGENCE IN INTENSITY, AND RECOMPOSITION TOWARD TRADE	107
Abdesh Dinara FROM A WOMEN'S LEAD TO A MEN'S LEAD: THE GENDER GAP IN DIGITAL PAYMENTS IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024	116
Amangeldi Aidana DID SMALL BUSINESS IN KAZAKHSTAN REALLY SHRINK IN 2026? THE NEW TAX CODE, REGIONAL TAX RATES AND THE FALL IN THE NUMBER OF INDIVIDUAL ENTREPRENEURS	124

Cheng Che	
BUILD, BUY, OR PARTNER? ENTRY MODE SELECTION AND RISK GOVERNANCE FOR WIND POWER INVESTMENT IN KAZAKHSTAN	133
Wang Liang	
FROM CAPACITY EXPANSION TO SYSTEM VALUE: COORDINATED DEVELOPMENT OF RENEWABLE ENERGY, GRID FLEXIBILITY AND ENERGY STORAGE IN KAZAKHSTAN	146
Chen Jun	
CONSTRUCTION-PERIOD CREDIT RISK MANAGEMENT FOR LARGE INFRASTRUCTURE PROJECT FINANCE IN KAZAKHSTAN	159
Alniyaz Danial	
HOW MUCH FUEL DOES A HYBRID REALLY SAVE? A SAME-MODEL COMPARISON OF HYBRID AND GASOLINE CARS IN US FUEL ECONOMY	173

ТЕХНИКАЛЫҚ ҒЫЛЫМДАР - ТЕХНИЧЕСКИЕ НАУКИ - TECHNICAL SCIENCE

Nurmakhan Almira	
ADVANTAGES OF FIBER OPTIC COMMUNICATION OVER TRADITIONAL COPPER CABLES	180
Kozhabai Bereke	
SECURITY CHALLENGES IN IOT-BASED TELECOMMUNICATION NETWORKS	188
Azamatuly Yerassyl	
WHY AI-GENERATED CODE FAILS A BENCHMARK: A FAILURE ANALYSIS OF THREE OPEN LANGUAGE MODELS ON 30 HUMANEVAL+ TASKS	197
Nurzhan Sanzhar	
A FEW SHELLS, A FEW OPERATORS: WHERE LOW EARTH ORBIT IS CROWDED IN AN ALTITUDE-INCLINATION MAP OF 16,620 ACTIVE SATELLITES	206
Askarbek Nurislam	
CAN THE WEATHER TELL WHICH DAYS IN ASTANA WILL BE POLLUTED? CLASSIFYING PM2.5 EXCEEDANCE DAYS WITH MACHINE LEARNING AND CHECKING THE DATA AGAINST A GROUND MONITOR	212
Iliyas Arman	
SHOOTING ON THE MOVE: HOW MUCH AIM A ROBOT LOSES BY IGNORING ITS OWN VELOCITY, AND WHETHER A VIRTUAL-TARGET CORRECTION RECOVERS IT	223
Ануар Әлімжан Бахытжанұлы, Сексенбаева Айша Курманбековна	
СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТАНДАРТОВ И МЕТОДИК ОБУЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ: NICE FRAMEWORK, ECSF И ПРАКТИКО- ОРИЕНТИРОВАННЫЕ ТЕХНОЛОГИИ	229
Mukhamedzhan Yerganat	
GROWTH OF GITHUB DEVELOPER COMMUNITIES IN KAZAKHSTAN, UZBEKISTAN AND KYRGYZSTAN, 2020-2026: EVIDENCE FROM THE GITHUB INNOVATION GRAPH	239

ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР – МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ – INTERDISCIPLINARY SCIENCES

Albiyev Asser	
HOW MUCH OF AN ARTICLE COUNT IS HUMAN WORK? DECOMPOSING THE CREATION OF FIVE WIKIPEDIAS, 2010–2025	251
Amangeldy Nurbolat	
HOW MUCH OF KAZAKHSTAN'S URBAN BUILT STOCK PREDATES ITS SEISMIC CODES, AND WHERE DOES THAT STOCK SIT RELATIVE TO HAZARD?	261

Torekhan Sabrina

WHERE ALMATY WALKS ITS OWN WAY: INFORMAL SHORTCUTS IN OPENSTREETMAP AND THE DETOURS THEY AVOID	267
---	------------

ӘЛЕУМЕТТІК ЖӘНЕ ГУМАНИТАРЛЫҚ ҒЫЛЫМДАР - СОЦИАЛЬНО-ГУМАНИТАРНЫЕ НАУКИ - SOCIAL SCIENCES AND HUMANITIES

Мұқатаева Айзат Амантайқызы, Стамбекова Жанат Байғазықызы, Слямханова Гүлім Болатқызы, Қызырбек Аяулым Ерболқызы

ЖАСАНДЫ ИНТЕЛЛЕКТ ДӘУІРІНДЕГІ ҒЫЛЫМИ ТАНЫМ: ПОППЕР, КУН ЖӘНЕ ЛАКАТОС ИДЕЯЛАРЫ ТҮРҒЫСЫНАН ТАЛДАУ	272
--	------------

Мұқатаева Айзат Амантайқызы, Сапарханова Мадина Төлегенқызы, Окенова Аружан Муратхановна, Тұрсынғазыева Ақерке Ерболатқызы

ШӘКӘРІМ ФИЛОСОФИЯСЫНДАҒЫ ЕРКІНДІК ЖӘНЕ МОРАЛЬДЫҚ ТАҢДАУ МӘСЕЛЕСІ	282
---	------------

Komekova Saida

DO UNHAPPY REVIEWERS MIX IN MORE RUSSIAN? STAR RATINGS AND KAZAKH-RUSSIAN CODE-SWITCHING IN 175,002 ONLINE REVIEWS	291
---	------------

Zhumagali Yerbol

DOES A BIGGER MODEL CLOSE THE GAP? KAZAKH-LANGUAGE ACCURACY ON KAZAKHSTAN-SPECIFIC VERSUS UNIVERSAL SCHOOL SUBJECTS IN SMALL OPEN-WEIGHT LANGUAGE MODELS	304
---	------------

Abdrakhmanova Madina

DOES A MEDAL BUY ATTENTION? WIKIPEDIA VIEWS OF KAZAKHSTAN'S ATHLETES BEFORE AND AFTER THE PARIS 2024 OLYMPIC	308
---	------------

ПЕДАГОГИКАЛЫҚ ЖӘНЕ ПСИХОЛОГИЯЛЫҚ ҒЫЛЫМДАР – ПЕДАГОГИЧЕСКИЕ И ПСИХОЛОГИЧЕСКИЕ НАУКИ – PEDAGOGICAL AND PSYCHOLOGICAL SCIENCES

Kaldarbek Mergen

FAMILY OR SCHOOL? SOCIOECONOMIC COMPOSITION AND MATHEMATICS ACHIEVEMENT IN KAZAKHSTAN, PISA 2022 AND 2025	316
--	------------

Zhenis Adilzhan

WHICH OLYMPIAD PROBLEMS DO STUDENTS TAKE TO THE BOARD? CHOICES AND DEFENCES IN THE 8TH-GRADE LEAGUE OF KARAGANDY MATHBATTLES 2026	323
--	------------

Базарбаева Айнура Бектуровна

ДЕТСКИЙ ФОЛЬКЛОР И ПОДВИЖНЫЕ ИГРЫ КАК ИНСТРУМЕНТЫ ФОРМИРОВАНИЯ ОСНОВ ПРАВОВОЙ КУЛЬТУРЫ У ДЕТЕЙ СТАРШЕГО ДОШКОЛЬНОГО ВОЗРАСТА	330
---	------------

БИОЛОГИЯ ҒЫЛЫМДАР - БИОЛОГИЧЕСКИЕ НАУКИ - BIOLOGICAL SCIENCES

Bazhenova Diana

GROWTH DYNAMICS OF STERLET (<i>Acipenser ruthenus</i> L.) REARED IN A RECIRCULATING AQUACULTURE SYSTEM	333
---	------------

ТУРИСТИЧЕСКИЕ И СПОРТИВНЫЕ НАУКИ - ТУРИСТІК ЖӘНЕ СПОРТ ҒЫЛЫМДАРЫ - TOURISM AND SPORTS SCIENCES

Төлеп Жазира Маратқызы

ГАСТРОНОМИЯЛЫҚ ФЕСТИВАЛЬДЕРДІҢ ТУРИСТІК ДЕСТИНАЦИЯ ТАРТЫМДЫЛЫҒЫН ҚАЛЫПТАСТЫРУДАҒЫ РӨЛІ МЕН ДАМУ ҮРДІСТЕРІ	340
--	------------

ЕСТЕСТВЕННЫЕ НАУКИ - ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ - NATURAL SCIENCES

Нургельдина Асель Ермаковна, Серикбаева Жанерке Алимжановна

ШЕКАРАСЫЗ ФИЗИКА: ӘР ЕЛДЕГІ ФИЗИКАЛЫҚ БІЛІМДІ САЛЫСТЫРУ 347