Dawid Jurczyński, MSc
WSB University in Dąbrowa Górnicza
e-mail: dawid.jurczynski@wsb.edu.pl
ORCID:0009-0005-7362-9779

Kamil Grzebień, MSc
WSB University in Dąbrowa Górnicza
e-mail: k.grzebien77@gmail.com
ORCID:0009-0007-1155-6070

DOI: 10.26410/SF_1/23/12

ESTIMATING THE POSSIBILITY OF A FIRE USING SATELLITE MISSIONS

Abstract

Space has been in the area of human interest for decades. Due to the multitude of satellite missions, we receive huge amounts of data every day that are used in many ways. The advantages of this data collection method include, among others, free access and the ability to record data without the need for physical sensors placed on Earth. This paper aims to present an attempt to use satellite missions to estimate the possibility of a fire in a specific area based on many environmental factors and the use of machine learning. The data used in this publication are based on satellite missions operating for the Copernicus program. The study area covers the territory of the Republic of Poland. The result of several models and Machine Learning methods were compared, including: under-sampling, over-sampling, catboost, randomforest. The research developed can serve to further develop work on a system that predicts the occurrence of a fire in an area. Through this research, it is possible to identify the most important relationships and features that correlate with the occurrence of fire.

Key words

Copernicus, possibility of a fire, satellites, machine learning

Introduction

The history of space exploration begins in 1957, when Soviet Union launched the world's first satellite "Sputnik 1"[1]. A year later, the United States of America sent into orbit its satellite "Explorer 1"[2]. The following years were marked by intense space exploration that resulted in a successful expedition to the Moon, the construction of an international space station or the launch of telescopes into space[3]. Today, space still fascinates humans, however space ventures reach further and further – like the Juice mission[4] that was intended to observe Jupiter. The amount of data we receive is very large[5]. One of the sources is the Copernicus program, which data will be used in this work.

Scientific research based on Sentinel satellite missions has been providing new information on topics related to ecology, agriculture and pollution. The applications of the data obtained are very wide and include issues such as: detection and monitoring of vegetation[6], assessments of water requirements for tomato crops[7], automated mosaicking imaginary[8], mapping Mediterranean seagrasses[9], crop water content of winter wheat revealed[10], road map extraction[11], atmospheric correction method[12], improving urban land cover classification[13], monitoring tidal flats boundaries[14], water quality retrieval[15] or mapping mangrove surface carbon stocks[16].

[1] https://aerospace.org/article/brief-history-space-exploration (access: 04.04.2023 r.).

[2] https://www.nasa.gov/mission_pages/explorer/explorer-overview.html (access: 04.04.2023 r.).

[3] https://education.nationalgeographic.org/resource/history-space-exploration/ (access: 04.04.2023 r.).

[4] https://www.esa.int/Science_Exploration/Space_Science/Juice (access: 22.04.2023 r.).

[5] https://latviaspace.gov.lv/en/news-events/dyk-cool-facts-about-the-eu-earth-observation-programme-copernicus/ (access: 22.04.2023 r.).

[6] M. Gašparović, D. Medak, I. Pilaš, L. Jurjević, I. Balenović, *Fusion of Sentinel-2 and planetscope imagery for vegetation detection and monitoring*, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci 2018, XLII-1, p. 155–160.

[7] A. Marta, G.Chirico, S. Bolognesi, M. Mancini, G. D'Urso, S. Orlandini, C. De Michele, F. Altobelli. *Integrating Sentinel-2 Imagery with AquaCrop for Dynamic Assessment of Tomato Water Requirements in Southern Italy*, "Agronomy" 2019, no. 7, p. 404.

[8] S. James, J. Schindler, J. R. Dymond, *Automated Mosaicking of Sentinel-2 Satellite Imagery*, "Remote Sensing" 2020, no. 22, p. 3680.

[9] D. Traganos, P. Reinartz. *Mapping Mediterranean seagrasses with Sentinel-2 imagery*, "Marine Pollution Bulletin" 2018, no 134, p. 197–209.

[10] D. Han, L. Shuaibing, D. Ying, X. Xinrui, F. Lingling, L. Lei, L. Zhenhong, Y. Hao, Y. Guijun, *Crop Water Content of Winter Wheat Revealed with Sentinel-1 and Sentinel-2 Imagery*. "Sensors" 2019, no. 18, 2019, p. 4013.

[11] C. Ayala, C. Aranda, M. Galar, *Towards fine-grained road maps extraction using sentinel-2 imagery*, "ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences" 2021, V-3-2021, p. 9–14.

[12] W. Su, Z. Mingzheng, J. Kunping, Z. Dehai, H. Jianxi, W. Pengxin, *Atmospheric Correction Method for Sentinel-2 Satellite Imagery*, "Acta Optica Sinica" 2018, no. 1.

[13] B. Hu, Y. Xu, X. Huang, Q. Cheng, Q. Ding, L. Bai, and Y. Li, *Improving Urban Land Cover Classification with Combined Use of Sentinel-2 and Sentinel-1 Imagery*, "ISPRS International Journal of Geo-Information" 2021, no. 8, 2021, p. 533.

[14] W. Yang, J. Sha, Z. Bao, J. Dong, X. Li, E. Shifaw, J. Tan, and T. H. Sodango, *Monitoring tidal flats boundaries through combining Sentinel-1 and Sentinel-2 imagery*, "Environmental Technology & Innovation"2021, p. 101401.

[15] M. Niroumand-Jadidi, F. Bovolo, M. Bresciani, P. Gege, C. Giardino, *Water Quality Retrieval from Landsat-9 (OLI-2) Imagery and Comparison to Sentinel-2*, "Remote Sensing" 2022, no. 18, 2022, p. 4596.

[16] A. Hakim, M. Kamal, S. Arjasakusuma, *Mapping mangrove surface carbon stocks using multisensor imagery in clungup mangrove conservation (CMC) malang regency*, "Jurnal Geografi" 2022, no. 2, 2022, p. 192.

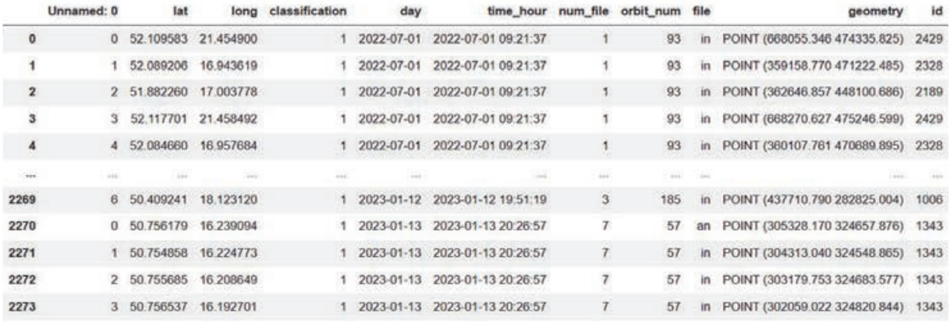
Acquisition and preparation of satellite data

The satellite data on which this work is based was obtained from the Copernicus Open Access Hub service[17]. It provides data from satellite missions such as Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5P. A Python-based library for simple data access is sentinelsat[18]. To be able to use it, you must have an account on the aforementioned website. The satellite data used in the program consists of elements such as temperature, dew point temperature, snow albedo, snow depth, soil wetness, surface pressure, total column water vapor, cloud fraction.

The first step is to define the area of our interest – in this case, it is the area of the country – Poland. The whole country is divided into squares – with a side of 10 kilometers. The whole is the basis for embedding specific data.

The next step is to group the points located in the area of a given square with a side of 10 kilometers and to draw the average value we are interested in. The average is also calculated from the measurements during one day. The whole thing is stored in the DataFrame object[19]. To obtain a graphical result on the map, a spatial join[20] process was carried out using the R-tree[21] algorithm.

Picture 1. An example geographic grid with geographic points



	Unnamed: 0	lat	long	classification	day	time_hour	num_file	orbit_num	file	geometry	id
0	0	52.109583	21.454900	1	2022-07-01	2022-07-01 09:21:37	1	93	in	POINT (668055.346 474335.825)	2429
1	1	52.089206	16.943619	1	2022-07-01	2022-07-01 09:21:37	1	93	in	POINT (359158.770 471222.485)	2328
2	2	51.882260	17.003778	1	2022-07-01	2022-07-01 09:21:37	1	93	in	POINT (362646.857 448100.686)	2189
3	3	52.117701	21.458492	1	2022-07-01	2022-07-01 09:21:37	1	93	in	POINT (668270.627 475246.599)	2429
4	4	52.084660	16.957684	1	2022-07-01	2022-07-01 09:21:37	1	93	in	POINT (360107.761 470689.895)	2328
...	...	...	...	...	...	...	...	...	...	...	...
2269	6	50.409241	18.123120	1	2023-01-12	2023-01-12 19:51:19	3	185	in	POINT (437710.790 282825.004)	1006
2270	0	50.756179	16.239094	1	2023-01-13	2023-01-13 20:26:57	7	57	an	POINT (305328.170 324657.876)	1343
2271	1	50.754858	16.224773	1	2023-01-13	2023-01-13 20:26:57	7	57	in	POINT (304313.040 324548.865)	1343
2272	2	50.755685	16.208649	1	2023-01-13	2023-01-13 20:26:57	7	57	in	POINT (303179.753 324683.577)	1343
2273	3	50.756537	16.192701	1	2023-01-13	2023-01-13 20:26:57	7	57	in	POINT (302059.022 324820.844)	1343

The next step is to overlay topographic objects from the BDOT10k[22] database on the prepared grid. These are data such as, among others: surface water, buildings, forest and wooded area, shrub vegetation, permanent crops, grassy vegetation and agricultural crops, area under roads, railroads and airports, unused land, yard, landfill, excavation and dumping ground, remaining undeveloped area. The purpose of this operation is to determine the number of square meters of a given terrain type in a particular place, and then determine the

[17] https://scihub.copernicus.eu/ (access: 11.04.2023 r.).

[18] https://sentinelsat.readthedocs.io/en/stable/index.html (access: 11.04.2023 r.).

[19] https://geopandas.org/en/stable/docs/reference/api/geopandas.sjoin.html (access: 13.04.2023 r.).

[20] https://geopandas.org/en/stable/gallery/spatial_joins.html (access: 23.05.2023 r.).

[21] https://geoffboeing.com/2016/10/r-tree-spatial-index-python/ (access: 13.04.2023 r.).

[22] https://www.geoportal.gov.pl/dane/baza-danych-obiektow-topograficznych-bdot (access: 13.04.2023 r.).

percent value. The filling of the known area with all collected data is shown in figure 2. The legend shows the percent value.

Picture 2. The percentage of known area data

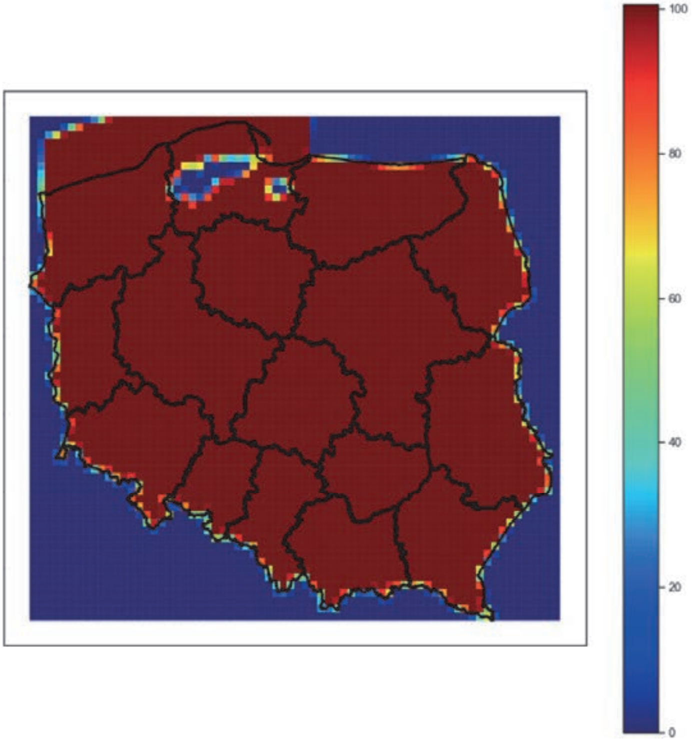
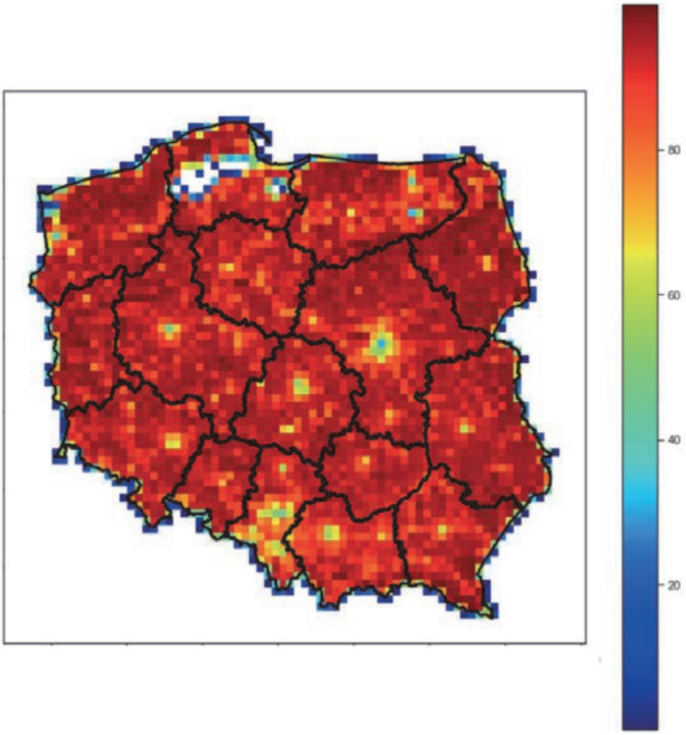


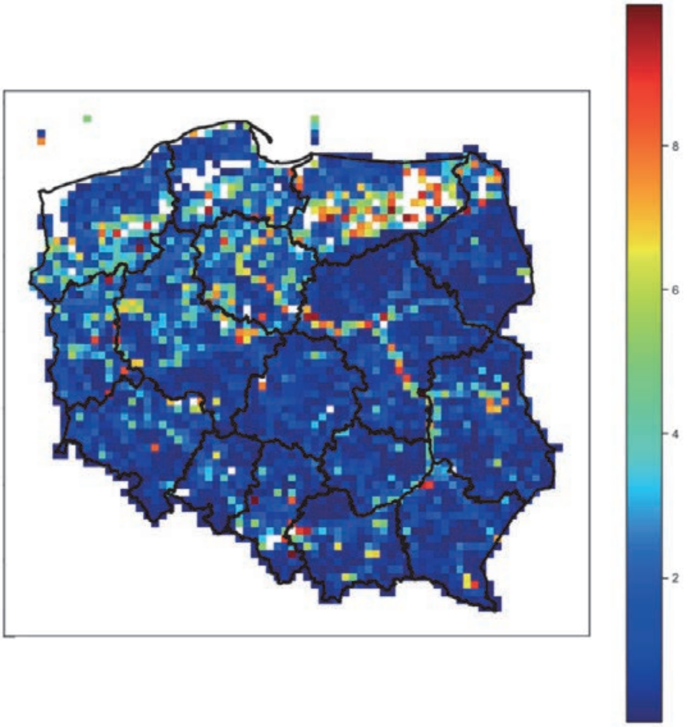
Figure 3 shows the percentage of vegetation in a given area. As in the previous case, its percent value was presented in appropriate color.

Picture 3. The percentage of vegetation in the area



The last graphical example is figure 4 – presenting the percentage of water in a given area. In this case, the value was limited to 10% of the given area, because otherwise the water recorded in the Baltic Sea would distort the graphical representation. In places in Poland where the area is shown in white, the percentage of water surface is greater than 10%.

Picture 4. The percentage of water in the area



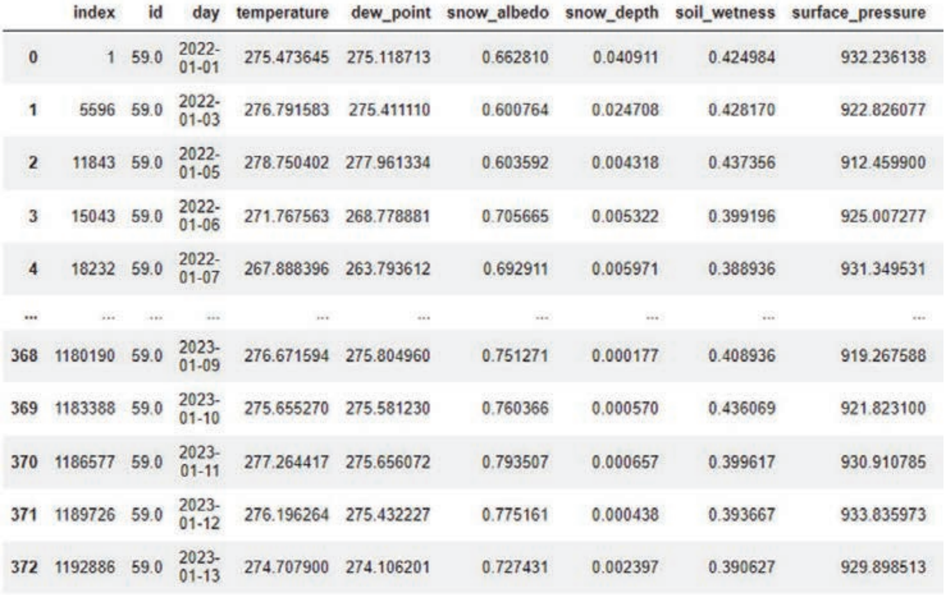
Using the data of the Sentinel-3[23] satellite mission, we take information recorded on Earth that later forms features included in the data model. They are listed at the beginning of this article. Moreover, they are also averaged across the individual cells on the map.

The next step is to add information about short-term trends in the time series for two values – soil moisture and temperature. This information is stored as a percentage change compared to the previous day and is used as a feature to build a machine-learning model.

Figure 5 shows a table section with geographical points with partially completed data. This is a Data Frame that will be populated with additional data in the next steps.

[23] https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-3/data-products (access: 22.04.2023 r.).

Picture 5. The section of the table with data in specific geographical points



	index	id	day	temperature	dew_point	snow_albedo	snow_depth	soil_wetness	surface_pressure
0	1	59.0	2022-01-01	275.473645	275.118713	0.662810	0.040911	0.424984	932.236138
1	5596	59.0	2022-01-03	276.791583	275.411110	0.600764	0.024708	0.428170	922.826077
2	11843	59.0	2022-01-05	278.750402	277.961334	0.603592	0.004318	0.437356	912.459900
3	15043	59.0	2022-01-06	271.767563	268.778881	0.705665	0.005322	0.399196	925.007277
4	18232	59.0	2022-01-07	267.888396	263.793612	0.692911	0.005971	0.388936	931.349531
...	...	...	...	...	...	...	...	...	...
368	1180190	59.0	2023-01-09	276.671594	275.804960	0.751271	0.000177	0.408936	919.267588
369	1183388	59.0	2023-01-10	275.655270	275.581230	0.760366	0.000570	0.436069	921.823100
370	1186577	59.0	2023-01-11	277.264417	275.656072	0.793507	0.000657	0.399617	930.910785
371	1189726	59.0	2023-01-12	276.196264	275.432227	0.775161	0.000438	0.393667	933.835973
372	1192886	59.0	2023-01-13	274.707900	274.106201	0.727431	0.002397	0.390627	929.898513

In the collected data, the occurrence of fire covers 0.1051% of the entire collection. With such a low percentage of fire occurrence, it would be possible to ignore this fact in the training set when dividing the data into 20% of the data in the training set and 80% of the data in the test set.

For this reason, the following division of data into training and test sets was adopted – the first step is to separate 20% of data with fires and 20% of data with no fire. The combined two sets were treated as training set.

Use of machine learning models in predicting dependencies

The first machine learning methods were *under_sampling*[24] and *over_sampling*[25] from the *imblearn*[26] libraries. The advantage of this approach is that these methods are applied to unbalanced data. Unfortunately, in the case under study no positive impact on the quality of prediction was observed.

Further tests were performed using the *catboost*[27] and *randomforest*[28] models. Better results were achieved by the catboost model. Figure 6 shows the two best sets of hyper-parameters.

[24] https://glemaitre.github.io/imbalanced-learn/api.html#module-imblearn.under_sampling (access: 18.04.2023 r.).

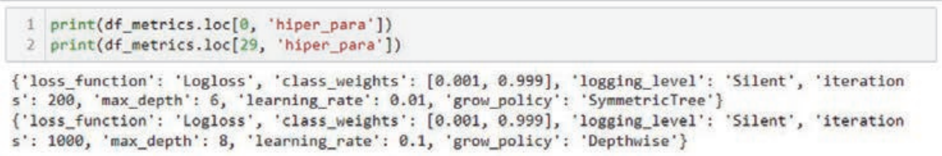
[25] https://glemaitre.github.io/imbalanced-learn/api.html#module-imblearn.over_sampling (access: 18.04.2023 r.).

[26] https://glemaitre.github.io/imbalanced-learn/index.html (access: 18.04.2023 r.).

[27] https://catboost.ai/en/docs/concepts/python-reference_catboost (access: 19.04.2023 r.).

[28] https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (access: 19.04.2023 r.).

Picture 6. The set of the best hyper-parameters for the catboost model

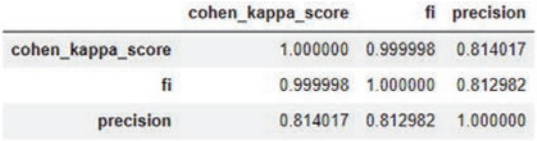


```
print(df_metrics.loc[0, 'hiper_para'])
print(df_metrics.loc[29, 'hiper_para'])
```

```
{'loss_function': 'Logloss', 'class_weights': [0.001, 0.999], 'logging_level': 'Silent', 'iteration
s': 200, 'max_depth': 6, 'learning_rate': 0.01, 'grow_policy': 'SymmetricTree'}
{'loss_function': 'Logloss', 'class_weights': [0.001, 0.999], 'logging_level': 'Silent', 'iteration
s': 1000, 'max_depth': 8, 'learning_rate': 0.1, 'grow_policy': 'Depthwise'}
```

Figure 7 shows the correlations between the rating metrics. The best metric turned out to be the *cohen_kappa_score*[29] metric. It gave the best results for the dataset used.

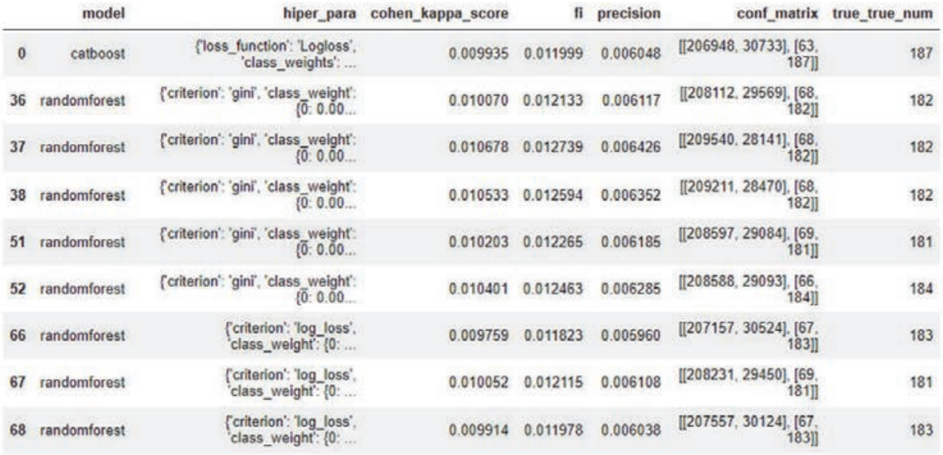
Picture 7. Correlations of metrics in the dataset



	cohen_kappa_score	fi	precision
cohen_kappa_score	1.000000	0.999998	0.814017
fi	0.999998	1.000000	0.812982
precision	0.814017	0.812982	1.000000

Figure 8 shows the confusion matrix with given metrics on the selected dataset. The goal is to select the best set of parameters to use in subsequent steps.

Picture 8. Selected metrics with error matrix



	model	hiper_para	cohen_kappa_score	fi	precision	conf_matrix	true_true_num
0	catboost	{'loss_function': 'Logloss', 'class_weights': ...	0.009935	0.011999	0.006048	[[206948, 30733], [63, 187]]	187
36	randomforest	{'criterion': 'gini', 'class_weight': {0: 0.00...	0.010070	0.012133	0.006117	[[208112, 29569], [68, 182]]	182
37	randomforest	{'criterion': 'gini', 'class_weight': {0: 0.00...	0.010678	0.012739	0.006426	[[209540, 28141], [68, 182]]	182
38	randomforest	{'criterion': 'gini', 'class_weight': {0: 0.00...	0.010533	0.012594	0.006352	[[209211, 28470], [68, 182]]	182
51	randomforest	{'criterion': 'gini', 'class_weight': {0: 0.00...	0.010203	0.012265	0.006185	[[208597, 29084], [69, 181]]	181
52	randomforest	{'criterion': 'gini', 'class_weight': {0: 0.00...	0.010401	0.012463	0.006285	[[208588, 29093], [66, 184]]	184
66	randomforest	{'criterion': 'log_loss', 'class_weight': {0: ...	0.009759	0.011823	0.005960	[[207157, 30524], [67, 183]]	183
67	randomforest	{'criterion': 'log_loss', 'class_weight': {0: ...	0.010052	0.012115	0.006108	[[208231, 29450], [69, 181]]	181
68	randomforest	{'criterion': 'log_loss', 'class_weight': {0: ...	0.009914	0.011978	0.006038	[[207557, 30124], [67, 183]]	183

The best score for the cohen_kappa_score metric is represented by the following parameters:

{'loss_function': 'Logloss', 'class_weights': [0.001, 0.999], 'logging_level': 'Silent', 'iterations': 1000, 'max_depth': 8, 'learning_rate': 0.1, 'grow_policy': 'Depthwise'}

[29] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html (access: 19.04.2023 r.).

The maps of Poland presented in the further part of the work are based on these hyper-parameters.

Figure 9 shows the probability of being classified into a given class. The closer the value is to *0*, the conclusion is that there is no fire. The closer the value is to *1*, it is more likely that a fire will occur. The limit value is *0.5*. Each point from the test base has its own probability presented.

Picture 9. The probability of being classified into a given class for the entire test set

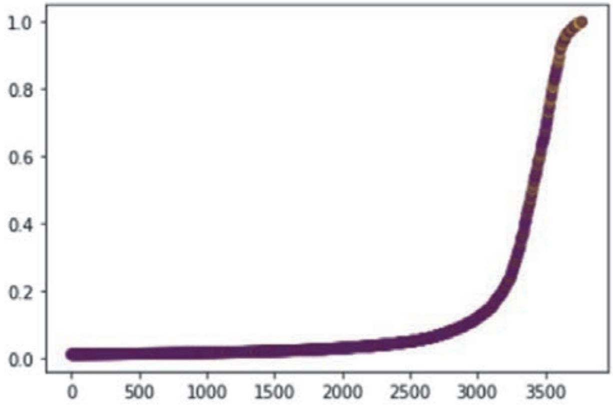


Figure 10 shows the probability of being classified into a given class. The difference from the previous graph is that only situations where a fire actually occurred were used.

Picture 10. The probability of classifying being classified into a given class for all situations where a fire actually occurred

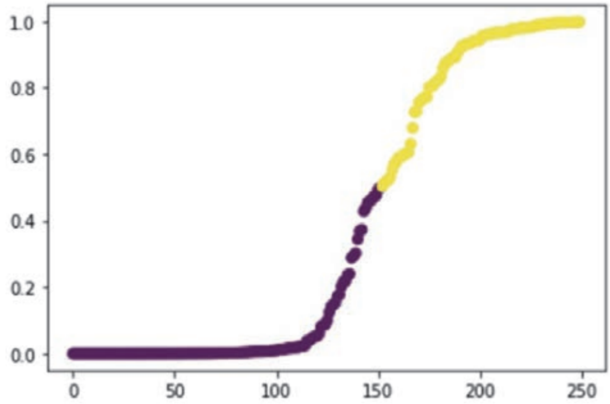
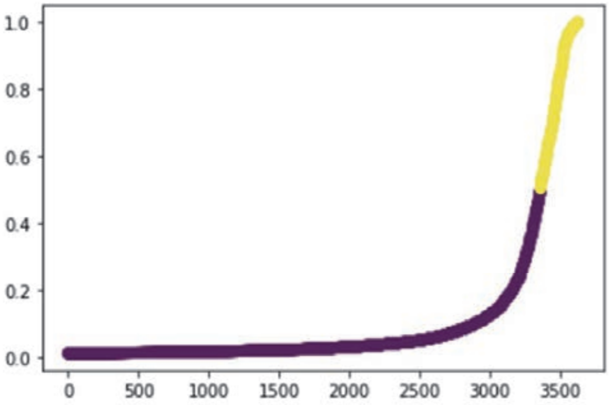


Figure 11 shows the opposite situation – the probability of classification to being classified into a given class where no fires occurred in a given data set.

Picture 11. Probability being classified into a given class for all situations where a fire did not occur



The next step was to slightly modify the probability of fire occurrence – in addition to the unchanged data set, a variant was also prepared in which, with a probability of fire occurrence greater than 1%, it was treated as a fact of fire occurrence. The unchanged variant shows the confusion matrix on the right, while the modified probability variant shows the confusion matrix on the left. These data are shown in figure 12. For the unmodified variant, the cohen_kappa_score value is: 0.3178484063078011.

Picture 12. Confusion matrix for the modified variant (left) and the unmodified variant (right)

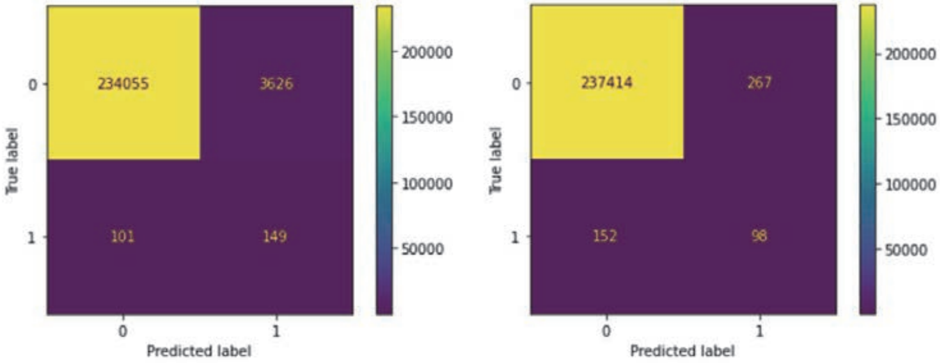
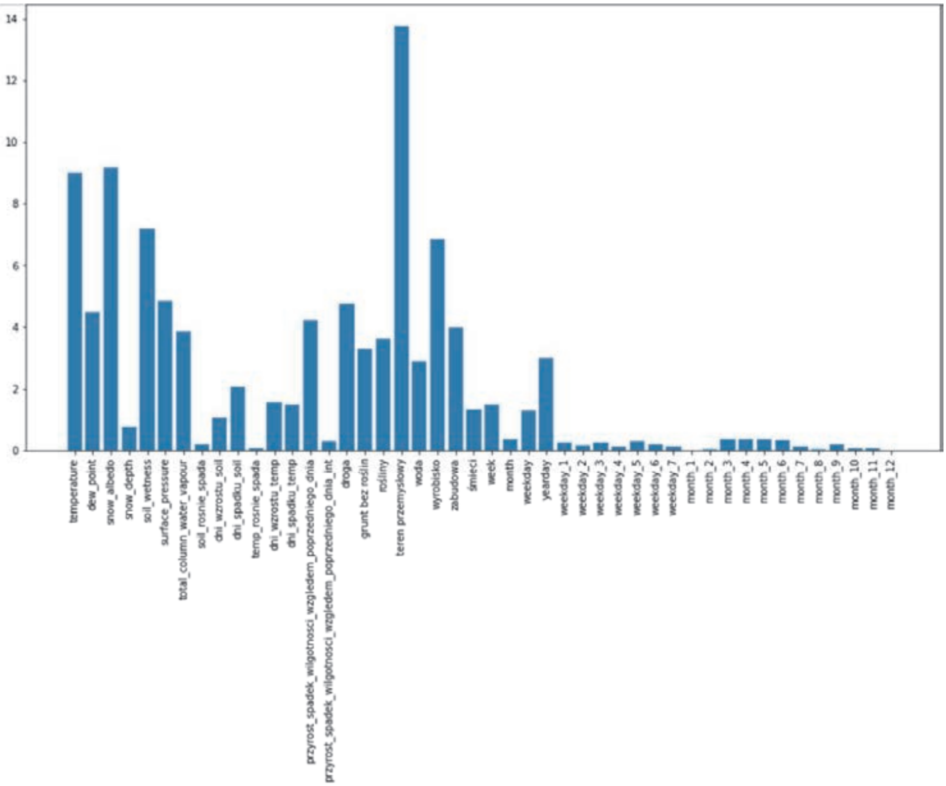


Figure 13 shows the significance of features in the prepared model. The height of the bar indicates how much the feature affects the final results.

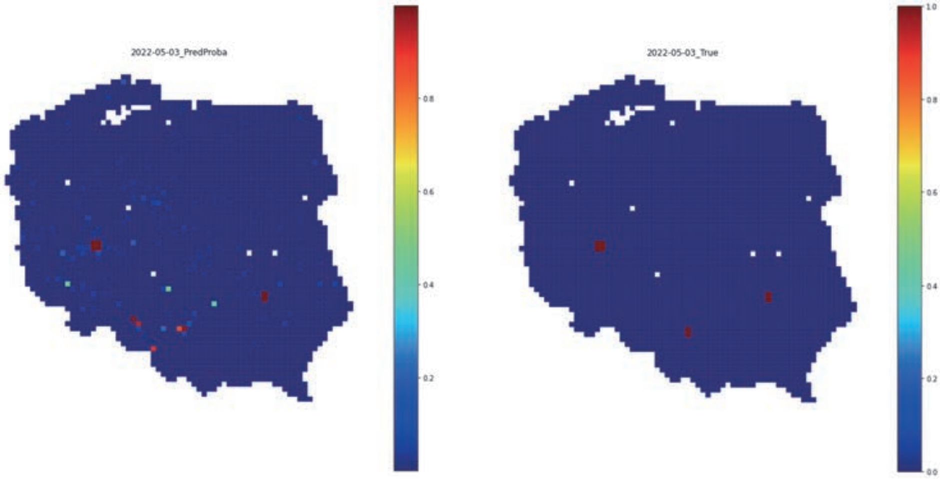
Picture 13. Significance of features in the prepared model



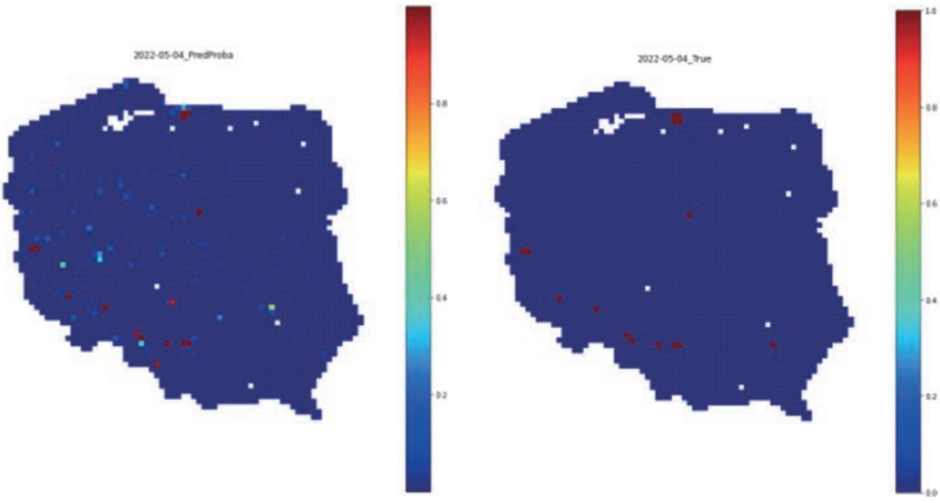
Summary

The final part of the work is to show the results in a graphical form – they show maps of Poland with the probability of fire occurrence and the actual state. This is presented in pictures number 14, 15, 16, 17 and 18. Each of them shows another day. The image on the left shows an attempt to predict fires, while the image on the right shows the actual occurrence of a fire. The color scale has values from 0 to 1. In predicted situations, different shades of colors represent the percentage probability of fire, while in the map representing real situations there are 2 possibilities – fire or no fire.

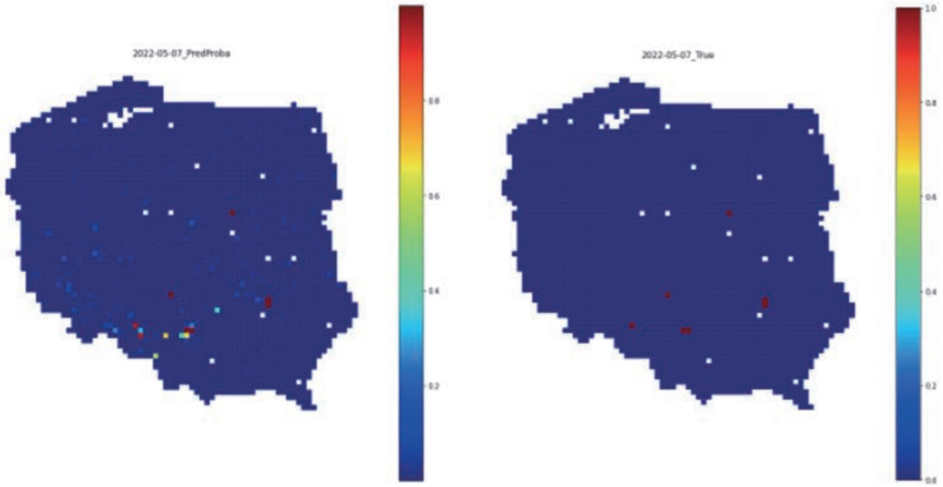
Picture 14. The probability of fire and the actual state on 03.05.2022



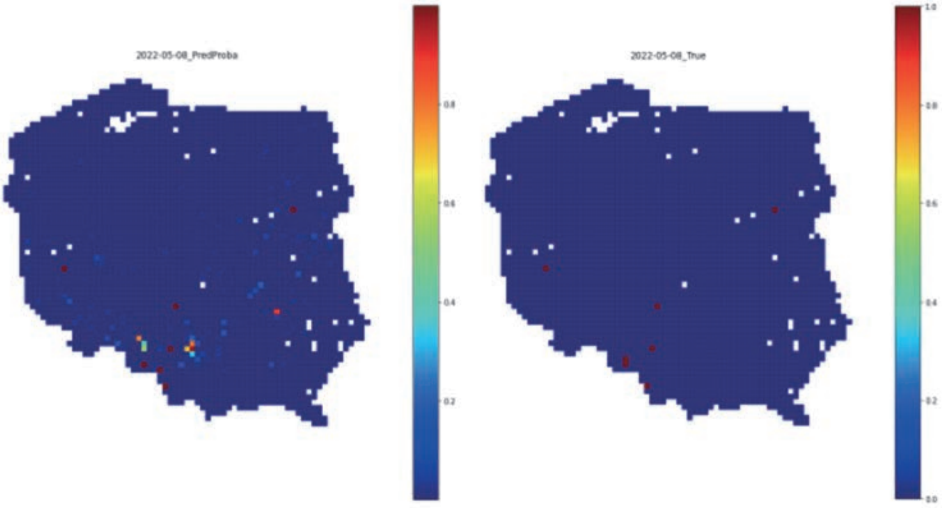
Picture 15. The probability of fire and the actual state on 04.05.2022



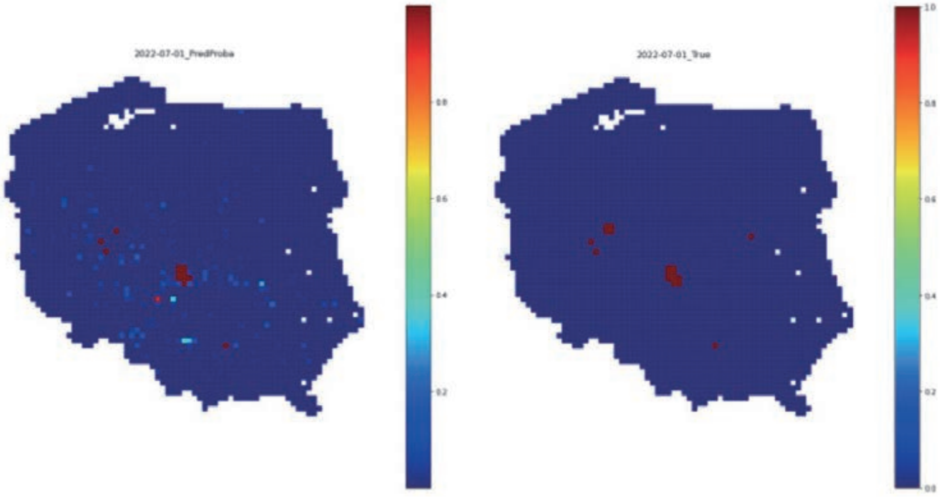
Picture 16. The probability of fire and the actual state on 07.05.2022



Picture 17. The probability of fire and the actual state on 08.05.2022



Picture 18. The probability of fire and the actual state on 01.07.2022



Based on the Confusion matrix[30] in figure 12, our model provides two possibilities with characteristics. In the unmodified variant, the fire number 98 was correctly found (True – Positive), the fire number 237414 was correctly found (True – Negative). The value of the False – Positive result is 267, while the False – Negative result is 152. In the modified variant, the values are as follows: True – Positive (predicted and actual fire) is 149, True – Negative (unforeseen fire and no actual fire) is 234055, False – Positive is 3626 False – Negative is 101.

To expand the tool and increase its accuracy, you can do the following:

- adding additional features to the data model,
- testing other machine learning methods,
- reducing the size of individual meshes of the grid (in the present case it is a square of 10 x 10 kilometers) to obtain a more accurate characterization of a given area.

Bibliography

Hakim A., Kamal M., Arjasakusuma S., *Mapping mangrove surface carbon stocks using multisensor imagery in clungup mangrove conservation (CMC) malang regency.* “Jurnal Geografi” 2022, no. 2, 2022.

Marta A., Chirico G., Bolognesi S., Mancini M., D’Urso G., Orlandini S., De Michele C., Altobelli F., *Integrating Sentinel-2 Imagery with AquaCrop for Dynamic Assessment of Tomato Water Requirements in Southern Italy*, “Agronomy” 2019, no. 7, 2019.

Hu B., Xu Y., Huang X., Cheng Q., Ding Q., Bai L., Li Y., *Improving Urban Land Cover Classification with Combined Use of Sentinel-2 and Sentinel-1 Imagery*, “ISPRS International Journal of Geo-Information” 2021, no. 8, 2021.

Ayala C., Aranda C., Galar M., *Towards fine-grained road maps extraction using sentinel-2 imagery*, “ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences” 2021, V-3-2021.

[30] https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (access: 22.04.2023 r.).

Han D., Shuaibing L., Ying D., Xinrui X., Lingling F., Lei L., Zhenhong L., Hao Y., Guijun Y., *Crop Water Content of Winter Wheat Revealed with Sentinel-1 and Sentinel-2 Imagery*, "Sensors" 2019, no. 18, 2019.

Traganos D., Reinartz P., *Mapping Mediterranean seagrasses with Sentinel-2 imagery*, "Marine Pollution Bulletin" 2018, 134.

Gašparović M., Medak D., Pilaš I., Jurjević L., Balenović I., *Fusion of Sentinel-2 and planetscope imagery for vegetation detection and monitoring*, "Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci" 2018, XLII-1.

Niroumand-Jadidi M., Bovolo F., Bresciani M., Gege P., Giardino C., *Water Quality Retrieval from Landsat-9 (OLI-2) Imagery and Comparison to Sentinel-2*, "Remote Sensing" 2022, no. 18, 2022.

James S., Schindler J., Dymond J. R., *Automated Mosaicking of Sentinel-2 Satellite Imagery*, "Remote Sensing"2020, no. 22.

Su W., Mingzheng Z., Kunping J., Dehai Z., Jianxi H., Pengxin W., *Atmospheric Correction Method for Sentinel-2. Satellite Imagery*, "Acta Optica Sinica" 2018, no. 1.

Yang W., Sha J., Bao Z., Dong J., Li X., Shifaw E., Tan J., Sodango T. H., *Monitoring tidal flats boundaries through combining Sentinel-1 and Sentinel-2 imagery*, "Environmental Technology & Innovation" 2021.

https://aerospace.org/article/brief-history-space-exploration.

https://www.nasa.gov/mission_pages/explorer/explorer-overview.html.

https://education.nationalgeographic.org/resource/history-space-exploration/.

https://www.esa.int/Science_Exploration/Space_Science/Juice.

https://latviaspace.gov.lv/en/news-events/dyk-cool-facts-about-the-eu-earth-observation-programme-copernicus/.

https://scihub.copernicus.eu/.

https://sentinelsat.readthedocs.io/en/stable/index.html.

https://geopandas.org/en/stable/docs/reference/api/geopandas.sjoin.html.

https://geopandas.org/en/stable/gallery/spatial_joins.html.

https://geoffboeing.com/2016/10/r-tree-spatial-index-python.

https://www.geoportal.gov.pl/dane/baza-danych-obiektow-topograficznych-bdot.

https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-3/data-products.

https://glemaitre.github.io/imbalanced-learn/api.html#module-imblearn.under_sampling.

https://glemaitre.github.io/imbalanced-learn/api.html#module-imblearn.over_sampling.

https://glemaitre.github.io/imbalanced-learn/index.html.

https://catboost.ai/en/docs/concepts/python-reference_catboost.

https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html.

https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5.

About the Authors

Dawid Jurczyński, Software Developer, Teaching Assistant. Fascinated by the opportunities presented by the use of satellite data and artificial intelligence.

Kamil Grzebień, Data Scientist, Data Analysis. Expert in ML modeling and data monetisation using visual and statistical methods on sensor, digital and IoT data.