Building a model for predicting fire dynamics

Lev Kene*, *and Roman* Kuzmich

Department of Business Informatics and Business Process Modeling, Siberian Federal University, 79 Svobodny Ave., Krasnoyarsk, 660041, Russia

Abstract. The paper compares known classification algorithms (logistic regression, k-nearest neighbors, support vector machine, stacking) on the problem of predicting the fire class. A feature of this problem is that it takes into account the specifics of the initial data at the exploratory analysis stage, i.e. before using the initial data by classification algorithms. First, at the exploratory analysis stage, it is necessary to solve the problem of selecting factors to predict the fire class. Second, check the initial data for gaps and outliers. Third, normalize the data in order to prepare them for building a model for predicting fire dynamics. Experimental studies are conducted on data on fires in the Krasnoyarsk Krai from 2010 to 2020.

1 Introduction

The Krasnoyarsk Krai, with its vast forests and diverse ecosystems, faces a serious problem of forest fires, which poses a threat to both the environment and the population. The increase in the frequency and intensity of fires in the region, partly due to climate change and anthropogenic activities, underscores the need to develop effective tools for monitoring and predicting the dynamics of fire disasters. The processing of data collected by departmental services using modern data analysis methods, including machine learning, will make it possible to predict the behavior of fires and assess the probability of their occurrence. This helps to better manage risks and optimize resource allocation for dealing with fire disasters. Therefore, the study of fire dynamics in the Krasnoyarsk Krai is important for developing strategies for prevention and risk management, contributing to the preservation of ecosystems and protecting the life and health of people in the region.

Information about fires is particularly valuable for public services, because statistics make it possible to plan resources more flexibly. Targeted planning of resources and countermeasures will reduce the damage: only for the period from 2018 to 2023, the total damage from forest fires amounted to 70 billion rubles. There are many information systems specialized in collecting and processing data on natural and anthropogenic fires. However, local decision makers often do not use the information collected in such systems [1], which is a big omission.

In this study, we analyze specific data on fires in the Krasnoyarsk Krai and then build a machine learning model to predict the dynamics of fires. The specifics of data analysis lie in the specifics of collecting data on emergency situations. The basis of technical monitoring is formed by terrestrial and aerospace instruments belonging to the relevant state bodies and

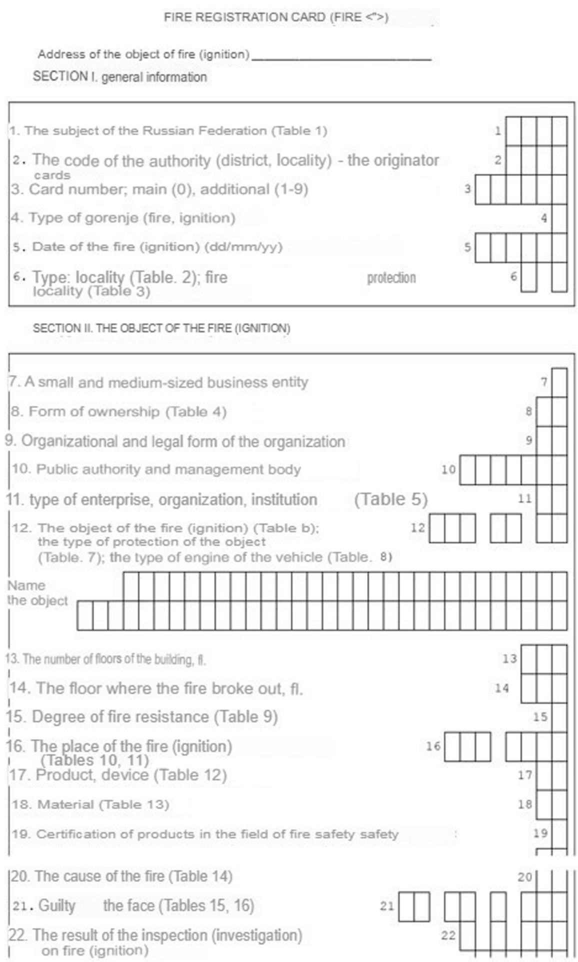
*Corresponding author: lev_kene@mail.ru

organizations, as well as enterprises. They act in accordance with their responsibilities and areas of responsibility, so the methods of collecting information about natural disasters may vary, despite the existing standard of data collection. Also, due to differences in collection methods, there may be data omissions, filling errors, and statistical outliers. All these features must be taken into account in the analysis.

Official statistical accounting is subject to all fires that involve legal entities and individual entrepreneurs licensed by the Ministry of Emergency Situations of Russia for the right to carry out fire extinguishing activities in localities, industrial facilities and infrastructure, fire protection units that are not licensees, as well as fires, in the liquidation of which fire departments and licensees did not participate, but information about which was received from individuals and legal entities [2].

Official statistical accounting of fires and their consequences in the Russian Federation is carried out by the federal fire Service directly and through the relevant structural divisions of the territorial bodies of the Ministry of Emergency Situations of Russia, which are responsible for organizing and implementing federal state fire supervision [2].

The data that is collected by departments is recorded first of all in the “fire accounting card”, also called “the Electronic Fire Accounting Card and Its Consequences (EFAC)”, it is shown in Figure 1.


FIRE REGISTRATION CARD (FIRE <">)

Address of the object of fire (ignition) ____________

SECTION I. general information

1. The subject of the Russian Federation (Table 1) 1
2. The code of the authority (district, locality) - the originator cards 2
3. Card number; main (0), additional (1-9) 3
4. Type of gorenje (fire, ignition) 4
5. Date of the fire (ignition) (dd/mm/yy) 5
6. Type: locality (Table. 2); fire protection locality (Table 3) 6

SECTION II. THE OBJECT OF THE FIRE (IGNITION)

7. A small and medium-sized business entity 7
8. Form of ownership (Table 4) 8
9. Organizational and legal form of the organization 9
10. Public authority and management body 10
11. type of enterprise, organization, institution (Table 5) 11
12. The object of the fire (ignition) (Table b); the type of protection of the object (Table. 7); the type of engine of the vehicle (Table. 8) 12

Name the object

13. The number of floors of the building, fl. 13
14. The floor where the fire broke out, fl. 14
15. Degree of fire resistance (Table 9) 15
16. The place of the fire (ignition) (Tables 10, 11) 16
17. Product, device (Table 12) 17
18. Material (Table 13) 18
19. Certification of products in the field of fire safety safety 19
20. The cause of the fire (Table 14) 20
21. Guilty the face (Tables 15, 16) 21
22. The result of the inspection (investigation) on fire (ignition) 22

Fig. 1. Fire accounting card.

Prior to 2018, the fire registration card was filled in by an official of the territorial department (department, inspection) of the State Fire Supervision Body (SPN) of the Main Directorate of the Ministry of Emergency Situations of Russia for the subject of the Russian Federation or the SPN body of a special or military division of the federal fire Service of the State Fire Service of the State Fire Service (FPS SBS) that supervised fires. Ministry of Emergency Situations of Russia Order No. 727 dated December 26, 2014 "On improving the formation of electronic databases for fire accounting and their consequences" regulates the process of filling in and passing fire accounting system cards [3].

The left part of the card contains a text part containing the names of code fields that are filled in with text or quantitative values. The right part of the card, also known as the code, is intended for automatic processing and is filled in the same order as the fire accounting card. The fire accounting card must be written or filled in with blue or black ink on white A4 sheets of paper in accordance with the established form [3].

Thus, the analysis of data on natural disasters is a complex and multifaceted task that requires deep knowledge in the field of mathematics, statistics, information technology, and also requires expert evaluation by specialists of the Ministry of Emergency Situations, environmentalists, firefighters, and other experts. Effective use of data on natural disasters can significantly improve our ability to prevent their occurrence, minimize their consequences, and ensure the safety of people and the environment. This is exactly what machine learning can help everyone with.

2 Methods and materials

In the hands of researchers were data from open sources of the Ministry of Emergency Situations of Russia. They were divided into 11 files with the extension ".csv". 11 files are data for each year from 2010 to 2020 inclusive. There are 204 parameters in this set, and 598950 rows. The parameters contain 200 factors, from f1 to f199. At the time of starting work, it was not known what these factors were, so we found special reference books and cards for deciphering the dataset fields in order to figure out what data each factor contains. The type and form of data recording is shown in Figure 2.

	A	B	C	D	E	F	G	H
28	F25	NULL	NULL	P19			P19	Date of the last detailed examination
29	F26	2	3	P20	4	1	P20	Distance to the fire station, km
30	F27	0	0	P21_1	1	0	P21_1	People died: total
31	F28	0	0	P21_2	0	0	P21_2	Including children
32	F29	0	0	P21_3	0	0	P21_3	of software workers
33	F30	0	0	P22_1	0	0	P22_1	Injured: total
34	F31	0	0	P22_2	0	0	P22_2	Including children
35	F32	0	0	P22_3	0	0	P22_3	of software workers
36	F33	0	0	P23_1	0	0	P23_1	Conditions that contributed to injury
37	F34	0	0	P23_2	0	0		
38	F35	0	0	P23_3	0	0		
39	F36	0	0	P24	0	0	P24	Direct damage
40	F37	0	0	P25	0	0	P25	By fixed assets
41	F38	0	0	P26	0	0	P26	By working capital
42	F39	0	0	P27_1	1	0	P27_1	DESTROYED: Buildings
43	F40	0	0	P27_2	0	0	P27_2	Units
44	F41	1	0	P28_1	0	0	P28_1	Residential apartments
45	F42	0	0	P28_2	0	0	P28_2	Rooms, units
46	F43	75	0	P29_1	24	0	P29_1	Floor area
47	F44	0	0	P29_2	0	1	P29_2	of Sq.m.
48	F45	0	0	P30_1	0	0	P30_1	Automotive equipment, units.
49	F46	0	0	P30_2	0	0	P30_2	Other equipment, units.
50	F47	0	0	P31_1	0	0	P31_1	Grain crops
51	F48	0	0	P31_2	0	0	P31_2	leguminous crops
52	F49	0	0	P32_1	0	0	P32_1	Feed
53	F50	0	0	P32_2	0	0	P32_2	Tons
54	F51	0	0	P33_1	0	0	P33_1	Technical cultures
55	F52	0	0	P33_2	0	0	P33_2	Tons
56	F53	0	0	P34	0	0	P34	Large cattle
57	F54	0	0	P35	0	0	P35	Small cattle
58	F55	0	0	P36	0	0	P36	Birds , pieces
59	F56	0	0	P37	0	0	P37	SAVED: people
60	F57	0	0	P38_1	0	0	P38_1	Livestock: large
61	F58	0	0	P38_2	0	0	P38_2	Livestock: small
62	F59	0	0	P39	0	0	P39	Automotive and other equipment
63	F60	0	0	P40	0	0	P40	Material values, RUB
64	F61	5	16	P41_1	2	7	P41_1	Detection time, hour.
65	F62	34	48	P41_2	45	15	P41_2	Detection time, min.
66	F63	5	16	P42_1	3	7	P42_1	Message time, hour.
67	F64	35	49	P42_2	6	23	P42_2	Message time, min.
68	F65	5	16	P43_1	3	7	P43_1	Arrival time of 1 fire department, hour
69	F66	42	53	P43_2	20	28	P43_2	Arrival time of 1 fire department, min.
70	F67	6	16	P44_1	3	7	P44_1	Localization time, hour
71	F68	2	55	P44_2	37	36	P44_2	Localization time, min
72	F69	6	16	P45_1	5	7	P45_1	Liquidation time, hour
73	F70	26	56	P45_2	21	58	P45_2	Elimination time, min
74	F71	03.01.2018 0:00	01.01.2018 0:00	P46	06-фев-00	11-фев-00	P46	Date of liquidation

Fig. 2. Reference guide for decoding dataset fields.

However, not all factors are informative for conducting fire analysis. Previously, this dataset was already studied by other researchers from the Reshetnev Siberian State University, and it was noted that the dataset does not contain measurements of temperature or fire propagation, so it is not possible to simulate the physical process of fire dynamics [4]. However, this data will help you create a fire class prediction model, a simplified representation of fire dynamics, so we need to prepare the data for analysis later for the machine learning model as well. Even so, we also need to select only the necessary factors for our model. An explanation of the factors that will be used for data analysis is presented in Table 1.

Table 1. Explanation of the dataset factors selected for analysis.

Factor	Explanation of the factor
f1	Region code
f5	Date of the fire
f12	The object of the fire
f14	Number of floors of the building
f15	The floor where the fire occurred
f16	Degree of fire resistance
f19	The cause of the fire
f26	Distance to the fire station, km
f27	People died: total
f29	Software workers died
f30	Injured: total
f32	Employees were injured by
f36	Direct damage
f61	Detection time, hour
f62	Detection time, min
f63	Message time, hour
f64	Message time, min
f65	Arrival time of 1 fire department, hour
f66	Arrival time of 1 fire department, min
f67	Localization time, hour
f68	Localization time, min
f69	Liquidation time, hour
f70	Elimination time, min
f71	Date of liquidation
f83	Technic
f88	Number of vehicles, units
f91	Trunks submitted
f94	Fire extinguishing agents
f100	Water sources
f148	Note
f149	The address of the object

Next, we'll start preprocessing the data. The first step in data preprocessing was to combine data from 11 years. Since each year contains about 600 thousand records, we need the necessary computing power, as well as a program that can merge big data. The Microsoft Power BI program was chosen to combine the data. Using the built-in query editor inside Power BI Desktop, called Power Query Editor, we can make the necessary changes to the data. Power Query Editor is an advanced business analysis tool designed for connecting to data sources and converting them [5]. In it, calculations were made and data was collected together for 11 years, from 2010 to 2020 inclusive, for a total of about 6 million rows with 204 columns. The calculations of about 15 minutes and were performed on a computer with the following characteristics: Intel Core i7-11800H 2.3 GHz, 32 GB of RAM, SSD, Windows 11 operating system.

Next, we selected the relevant parameters and renamed the columns to make them easier to understand. Since the dataset contains records for the whole of Russia, we selected data related to the Krasnoyarsk Krai, region code is 1104.

The next task is to process the address where the fire occurred, this parameter is significant. The data that was recorded in the "Object Address" is not standardized, so it needs to be processed, while in the rest of the data filled in the data set is already encoded. Somewhere the information is contained in full form, for example, "663400, Motyginsky district, Motygino village, Shosseynaya street, 74" – where there is an index, district, locality, street, house, and sometimes other information is indicated. Somewhere on the contrary, there is too little information, for example, it contains only information about the locality "city Achinsk". Since information about a locality is significant, it was decided to clear the data in the "Address of the object" from other information and leave only information about the locality.

As part of the analysis of fire data in the Krasnoyarsk Krai, we work with various types of information, including the location of fire centers, timestamps, and factors that detail the situation and nature of the fire. The data structure is presented in the form of a table, where each row corresponds to a separate fire incident, containing the fire ID, key timestamps regarding the fire object, and other necessary data about it.

The data is high-quality. During preprocessing, problems were found only in the "Object Address" and "Notes" columns. If the "Notes" column is filled in freely, then this freedom is not allowed for the "Object Address" column. If the address of an object is unknown, then at least it should contain data in the form of coordinates – meridian and parallel, but this is a controversial point. Due to the unified way of data collection, there is no manipulation in the data, which means that the data is reliable.

Now let's look at the descriptive statistics for the generated dataset for analysis. It is shown in Figure 3.

Key information from descriptive statistics is as follows:

- the largest number of records was recorded on April 27, 2017 – 413 fires.
- basically, 1-2 fire crews are sent to extinguish the fire;
- the largest number of entries belongs to the locality of Krasnoyarsk – 41763 entries.
- there is an anomaly in the calculated difference between the message and fire elimination, as well as the difference between the message and arrival (an error was found, there is incorrectly filled in data).

The source [6] describes in detail the steps and methods of exploration analysis. We will use some of them. First of all, it is necessary to plot histograms on the subject of studying the distribution of values, this will allow you to immediately study the data for the presence of dependencies. A scatter plot that can be built using Python and the Pandas plugin library will help Pandas you do this. Figure 4 shows the scattering matrix for all indicators.

We also check the data for gaps by calling the isna().sum() method, the result is shown in Figure 5.

	count	unique	top	freq	mean	min	25%	50%	75%	max	std
Region Code	158681.0	NaN	NaN	NaN	1104.0	1104.0	1104.0	1104.0	1104.0	1104.0	0.0
Date of Fire Occurrence	158681	3989	27 апреля 2017 г.	413	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fire Object	158681.0	NaN	NaN	NaN	332.238749	65.0	283.0	362.0	367.0	526.0	72.399497
Building Height	158681.0	NaN	NaN	NaN	0.647551	0.0	0.0	0.0	1.0	99.0	1.847637
Floor Where Fire Started	158681.0	NaN	NaN	NaN	4.018736	0.0	0.0	0.0	1.0	908.0	55.786237
Fire Resistance Class	158681.0	NaN	NaN	NaN	2.101109	0.0	0.0	0.0	5.0	6.0	2.540923
Cause of Fire	158681.0	NaN	NaN	NaN	12.967898	0.0	0.0	11.0	25.0	53.0	13.04693
Distance to Fire Station (km)	158681.0	NaN	NaN	NaN	7.290879	0.0	3.0	4.0	8.0	800.0	11.453092
Total Casualties	158681.0	NaN	NaN	NaN	0.017551	0.0	0.0	0.0	0.0	9.0	0.164772
Casualties among Fire Service Workers	158681.0	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Injuries	158681.0	NaN	NaN	NaN	0.016757	0.0	0.0	0.0	0.0	8.0	0.1565
Injuries among Fire Service Workers	158681.0	NaN	NaN	NaN	0.000101	0.0	0.0	0.0	0.0	3.0	0.011774
Direct Damage	158681.0	NaN	NaN	NaN	18592.095449	0.0	0.0	0.0	0.0	276730804.0	855106.906045
Detection Time (hour)	158681.0	NaN	NaN	NaN	13.455379	0.0	10.0	15.0	19.0	24.0	6.532893
Detection Time (minute)	158681.0	NaN	NaN	NaN	28.217695	0.0	13.0	28.0	44.0	59.0	17.633447
Report Time (hour)	158681.0	NaN	NaN	NaN	13.462204	0.0	10.0	15.0	19.0	24.0	6.545423
Report Time (minute)	158681.0	NaN	NaN	NaN	28.652157	0.0	13.0	29.0	44.0	59.0	17.5981
Arrival Time of 1st Fire Unit (hour)	158681.0	NaN	NaN	NaN	13.402733	0.0	9.0	15.0	19.0	24.0	6.664814
Arrival Time of 1st Fire Unit (minute)	158681.0	NaN	NaN	NaN	28.590222	0.0	13.0	28.0	44.0	59.0	17.595003
Containment Time (hour)	158681.0	NaN	NaN	NaN	13.076348	0.0	8.0	15.0	19.0	24.0	6.97004
Containment Time (minute)	158681.0	NaN	NaN	NaN	27.589724	0.0	12.0	27.0	43.0	59.0	17.902251
Elimination Time (hour)	158681.0	NaN	NaN	NaN	13.150031	0.0	8.0	15.0	19.0	24.0	6.990046
Elimination Time (minute)	158681.0	NaN	NaN	NaN	27.829028	0.0	12.0	28.0	43.0	59.0	17.896289
Date of Elimination	158185	3990	27 апреля 2017 г.	409	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Equipment	158681.0	NaN	NaN	NaN	11.167197	0.0	11.0	11.0	11.0	82.0	4.593406
Number of Equipment Units	158681.0	NaN	NaN	NaN	1.505127	0.0	1.0	1.0	2.0	51.0	1.06156
Hoses Deployed	158681.0	NaN	NaN	NaN	2.231174	0.0	2.0	2.0	2.0	9.0	1.414456
Number of Hoses	158681.0	NaN	NaN	NaN	1.144687	0.0	1.0	1.0	1.0	46.0	0.695493
Extinguishing Agents	158681.0	NaN	NaN	NaN	7.25349	0.0	0.0	11.0	11.0	19.0	5.446405
Water Sources	158681.0	NaN	NaN	NaN	4.007953	0.0	0.0	6.0	7.0	8.0	3.280028
Notes	158635	20168	сухая трава	25072	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Index	158681.0	NaN	NaN	NaN	2913403.41308	963.0	1099956.0	3217623.0	4507708.0	5406811.0	1796031.687888
Date and Time of Fire Detection	158681	NaN	NaN	NaN	2015-01-23 02:42:40.772367360	2010-01-01 00:00:00	2012-05-11 11:45:00	2014-12-08 07:21:00	2017-06-28 21:26:00	2020-12-02 23:38:00	NaN
Date and Time of Report	158681	NaN	NaN	NaN	2015-01-23 02:43:29.776721664	2010-01-01 00:00:00	2012-05-11 11:47:00	2014-12-08 07:22:00	2017-06-28 21:27:00	2020-12-02 23:39:00	NaN
Date and Time of Arrival of 1st Fire Unit	158681	NaN	NaN	NaN	2015-01-23 02:39:48.151700480	2010-01-01 00:00:00	2012-05-11 11:40:00	2014-12-08 07:28:00	2017-06-28 21:35:00	2020-12-02 23:44:00	NaN
Date and Time of Elimination	158681	151386	00:00	388	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Time Difference between Report and Arrival	158681.0	NaN	NaN	NaN	-3.69375	-1438.0	4.0	7.0	11.0	1404.0	131.693873
Locality	158036	20114	Красноярск	41763	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Time Difference between Report and Elimination	158681.0	NaN	NaN	NaN	-185526.768296	-63568020.0	0.0	0.0	60.0	437760.0	3315072.17089
Type of Fire	158681	2	простой	155902	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 3. Descriptive statistics of the dataset for analysis.

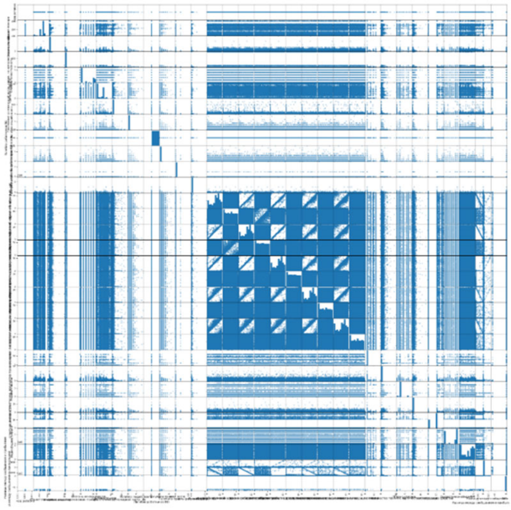


Fig. 4. The scattering matrix.

```
df.isna().sum()

Region Code                                      0
Date of Fire Occurrence                          0
Fire Object                                      0
Building Height                                  0
Floor Where Fire Started                         0
Fire Resistance Class                            0
Cause of Fire                                    0
Distance to Fire Station (km)                    0
Total Casualties                                 0
Casualties among Fire Service Workers            0
Total Injuries                                   0
Injuries among Fire Service Workers              0
Direct Damage                                    0
Detection Time (hour)                            0
Detection Time (minute)                          0
Report Time (hour)                               0
Report Time (minute)                             0
Arrival Time of 1st Fire Unit (hour)             0
Arrival Time of 1st Fire Unit (minute)           0
Containment Time (hour)                          0
Containment Time (minute)                        0
Elimination Time (hour)                          0
Elimination Time (minute)                        0
Date of Elimination                            496
Equipment                                        0
Number of Equipment Units                        0
Hoses Deployed                                   0
Number of Hoses                                  0
Extinguishing Agents                             0
Water Sources                                    0
Notes                                           46
Index                                            0
Date and Time of Fire Detection                  0
Date and Time of Report                          0
Date and Time of Arrival of 1st Fire Unit        0
Date and Time of Elimination                     0
Time Difference between Report and Arrival       0
Locality                                       645
Time Difference between Report and Elimination   0
Type of Fire                                     0
dtype: int64
```

Fig. 5. Checking for omissions.

As can be seen from the check, omissions are observed in the parameters “Date of liquidation”, “Note”, “Locality”. Lines with omissions should be deleted so as not to pollute the statistics, but you don't need to clean the lines in “Note”.

It is important to look at the data for strong outliers in the data. To do this, plot a box with a mustache, also called a boxplot, for each parameter and examine the columns for outliers. Figure 6 shows the box graphs with whiskers for each parameter.

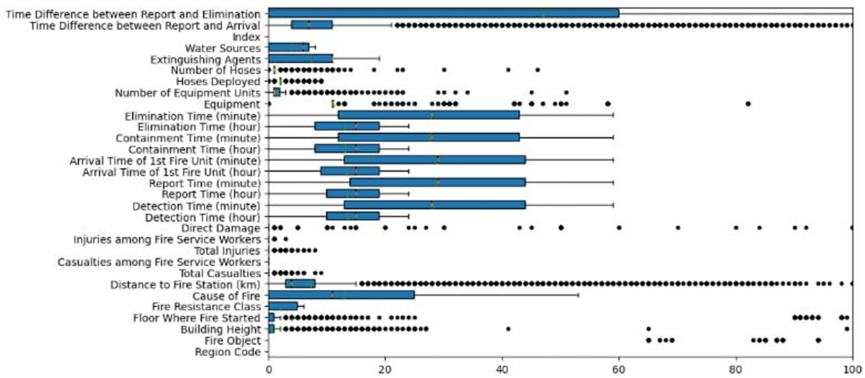


Fig. 6. Boxplot diagrams for dataset.

The data contains outliers in many indicators, but many outliers are significant, for example, by the factor "Distance to the fire station, km". Since data manipulation is excluded, outliers are natural and do not require deletion.

3 Results

In the course of studying the data and the ability to study the dynamics of fires, a classification task was selected. So, the goal of building a machine learning model is to predict how complex a fire will be – simple or difficult. A difficult fire is one that is extinguished for 6 or more hours. Since the forecast must be made at the stage of calling the fire department, not all data is known initially. In this regard, we need to create another dataset, which will become a training sample. Key parameters can be known immediately when a fire is reported, so we can make a management decision at the call stage – how many vehicles and/or fire departments to send to the site. The selected parameters for the training sample with interpretation are shown in Table 2.

Table 2. Explanation of the factors selected for building the model.

Factor	Explanation of the factor
f1	Date of the fire
f5	Type: locality
f12	of the fire protection of the locality
f14	The object of the fire
f15	Number of floors of the building
f16	The floor where the fire occurred
f19	Degree of fire resistance
f26	Place of origin
f27	Distance to the fire station, km
f29	Detection time, hour
f30	Detection time, min
f32	Message time, hour
f36	Message time, min
f61	Arrival time of 1 fire department, hour
f62	Arrival time of 1 fire department, min
f63	Liquidation time, hour

Factor	Explanation of the factor
f64	Elimination time, min
f65	Date of liquidation
f66	Technic
f67	The address of the object
f68	Localization time, min
f69	Liquidation time, hour
f70	Elimination time, min
f71	Date of liquidation
f83	Technic
f88	Number of vehicles, units
f91	Trunks submitted
f94	Fire extinguishing agents
f100	Water sources
f148	Note
f149	The address of the object

After preprocessing, the columns “Difference between arrival and liquidation” and the target variable “Type of fire”, “Date of fire occurrence” were formed, divided into “month of detection” and “day of detection”. Now in the new dataset, we need to delete the intermediate columns: “Date of fire occurrence”, “Detection time, min.”, “Message time, min.”, “Date and time of fire detection”, “Date and time of message”, “Date and time of liquidation”, “Difference between arrival and liquidation”, “Liquidation time, hour”, “Liquidation time, min”, “Liquidation date”, “Arrival time of 1 fire units, hour”, “Arrival time of 1 fire department, min.”, “Date and time of arrival of 1 fire department”.

The Jupyter Notebook environment was used for the design. Let's connect the libraries Numpy, Pandas, Matplotlib, Seaborn, which are the basis for data analysis in Python [7], and the sci-kit learn library, which contains the most popular machine learning models: classification, regression and clustering models, ensemble solutions, scaling methods, metrics. Next, we will upload a file with preprocessed data. When checking for passes, 105 passes were found in the column “Locality”, all these lines were deleted. Next, it was decided to train the model on a good sample with minimizing errors, so we will get rid of outliers in numerical columns using the interquartile range. After that, we will categorize the data.

After encoding categorical variables, it is necessary to update the data, as well as to scale – normalize the data. Normalization is the process of bringing data to a standard form or range in order to ensure their comparability and improve the performance of machine learning algorithms. Examples of normalization methods include min-max normalization, Z-normalization, and logarithmic transformation. Normalization is necessary to avoid problems with data scaling, improve the convergence of algorithms and increase the accuracy of forecasts, MinMax-normalization was used.

Now, to evaluate the quality of the model, you need to divide the data into training and test samples in the ratio 3: 2. The training sample is used to train the model, and the test sample is used to check its performance. It is also important to take into account the imbalance of classes – only 3% of the entire sample are difficult fires. To do this, we will use RandomOverSampler() to artificially increase the small class and balance the classes, but we will perform these transformations only for the training sample.

Logistic regression, the k-nearest neighbour method, the support vector machine (SVM) and stacking were chosen as models for solving the problem of predicting the type of fire.

Accuracy, classification report (precision, recall, f1-score, error matrix), ROC-curve, error matrix with heat map are output from statistics. First, a grid of 2-3 hyperparameters is set, changing which the search for the best model will be carried out. Training takes place with a grid search, the best model is highlighted. Next, this model is trained on a training sample and the accuracy of predictions and the classification report are output. A ROC-curve is constructed and an error matrix with a heat map is output.

In the Mariupolskiy Timofey research a fire classification model was built for the presence of victims, and the developed algorithm based on regression models such as gradient boosting, random forest, ridge and lasso regressions received an accuracy of 78% [8]. Whereas our algorithms turned out to be more successful.

4 Conclusion

When comparing the results, it was found that the most successful model is the SVM, based on the ROC-curve. Speaking of accuracy, the accuracy of the support vector machine was 83%, while the best accuracy was achieved by k-nearest neighbour and stacking methods – 97% and 96%, respectively. Since there is an imbalance of classes in the test sample, and the real situation is simulated, it is more correct to rely on metrics such as f1-measure and accuracy. Hence, the best is k-nearest neighbour model.

The desired result is more than 80% accurate, and it is observed in 3 out of 4 models, which means that machine learning methods are great for working with fire data, but presumably neural networks will be somewhat more efficient than machine learning – they will learn faster and give a better result.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2022-1121).

References

1. V.S. Rabdanov, I.A. Panfilov, E.D. Kolmykova, *Development of a Risk Management System for Combating Forest Fires*, in Proceedings of the All-Russian Scientific and Practical Conference with International Participation “Law, Economics, and Management: State, Problems, and Prospects”, August 22, 2024, Cheboksary, Russia (2024)
2. D. Brzezińska, Safety and Fire Technology, **42,** 141-149 (2016).
3. V.V. Pitsyk, A.B. Ichmelyan, Fires and emergencies: prevention, elimination **4,** 157-164 (2024). doi:10.25257/FE.2024.4.157-164
4. T.I. Mariupolsky, V.V. Nicheporchuk, D.V. Kustov, V.V. Rudenko, Scientific and Technical Bulletin of the Volga Region, **6**, 80 (2024)
5. D.G. Potemka, Applied Mathematics: Modern Problems of Mathematics, Informatics, and Modeling, **2**, 38 (2020)
6. D.A. Elsukov, Economy and Society **11-1(90),** 982-985 (2021).
7. O. I. Zakharova, I.Z. Kvachakhia, Bulletin of Science, **6(75)**, 1367 (2024)
8. T.I. Mariupolsky, I.A. Panfilov, Modeling Damage Based on Natural and Anthropogenic Disaster Data, in *Proceedings of the All-Russian Scientific and Practical Conference with International Participation “Law, Economics, and Management: State, Problems, and Prospects”,* August 22, 2024, Cheboksary, Russia (2024)