Classification of federal states of Brazil based on threat by the forest fires

Bojana Ivanovic Mijatov [1][0000-0002-1588-5396] and Vanja Mijatov [2][0000-0001-6298-2239]

[1] University of Novi Sad, Faculty of Technical sciences, 21000 Novi Sad, Serbia
bojana.ivanovic@uns.ac.rs
[2] University of Novi Sad, Faculty of Technical sciences, 21000 Novi Sad, Serbia
vanja.mijatov@uns.ac.rs

Abstract. Forest fires are causing great problems in Brazil where majority of the land is under the forests, with Amazon being the largest and most famous one, often referred to as the lungs of the world. A big step in solving the problem and thus reducing the damage caused by the forest fires would be if competent authorities could have an insight in which of the regions are the most threatened ones. This paper offers way to classify federal states of Brazil in different times of the year according to the level of vulnerability using several different methods for classification. Data set used data regarding socio-economical and temporal data and the idea was to see their influence on forest fires in federal states of Brazil during different periods of the year. Data included a lot of features, but the ones that attracted attention were the month and year in which the fire has occurred and the area of the federal state. Best results were obtained by using random forest algorithm for the classification problem. Some other algorithms like SVM were expected to give better results than they did, since they apply well on this kind of classification problems.

Keywords: Forest fires, Brazil, classification, random forest, SVM, ensemble algorithms.

1 Introduction

Every day on television and the Internet you can hear and read news about the fires that are raging around the planet and no part of the world is spared. Fires leave a lot of damage behind them, destructing both material properties and biodiversity. Even though the material property damage is more important to people nowadays and is measured in huge amount of money, the damage suffered by nature is much greater. Every year, vast of land is swallowed by forest burns and thus our planet loses the green cover that means life to all of us. All these fires in nature are a direct consequence of human action (ignition) or climate changes.

Of all the fires that happen in the world, these that happen in Brazil are the most worrying. This country is home for numerous different plant species. Brazil's most famous part is the Amazon rainforest, often referred to as the lungs of the world. In

addition to being a home for many plants and animals, the rainforest is the largest producer of oxygen on Earth. The forest fires that occur every year in this rainforest reduce the production of oxygen and make it difficult for the rainforest to recover and return to the previous state. The aim of this paper was to perform an exploratory analysis of data on fires in Brazil, which were collected for more than two decades and then use the conclusions of analysis to implement and improve classification of states by vulnerability in different periods in the year. This way we would have an insight into possible fires, their locations and that information could be further used for prevention of these fires or at least partial localization. Even though most of the papers address this problem as prediction problem, our idea was to observe it as a classification problem. Furthermore, our idea was to use socio-economical and temporal data and to see their influence on forest fires in federal states of Brazil during different months of the year.

In continuation of this paper, section Related work presents an overview of the relevant papers which were the basis for the proposed solution as well as their achievements. Section Data sets and preprocessing highlights which data sets were used, which transformations were used on the data and how data sets were combined into a final data set that was used. Exploratory data analysis is also discussed in this section. Methodology section gives an overview of methodologies that were used to solve the problem and at the very end of the paper, within the Conclusion, the paper is summarized, and some possible directions of research were highlighted.

2 Related work

During our research we took into consideration papers that observed forest fire problem as a prediction one, since we wanted to investigate how authors generally addressed these kinds of problems and which were the challenges they have encountered. In the paper [1], the authors from Indonesia addressed the problem of fire prediction, as well as the size of the areas affected by fires in different regions based on the data that belongs to the state institutions. Data which they used included number of fires for many years, and it was grouped by certain time intervals. Data was collected using semantic networks, that was provided by Linked open government data (LOGD) which enables querying and obtaining larger amount of publicly available data collected by the state institutions. Several approaches were applied on data: linear regression, backpropagation neural network and SVM algorithm. Authors have pointed out that the best results were obtained by using linear regression and that their approach is applicable to the other problems with similar data set, while also pointing out that the solution they came up with was not fully optimized for the given problem.

Authors of the paper [2] dealt with the idea of determining the probability of forest fires and the influence of weather factors on the occurrence of forest fires in Rokan Hilir district in Riau province in Indonesia. In their research, they used data describing the main hotspots and spread of the fires, weather data (maximum daily temperature, daily precipitation, and wind speed) and data on human activity factors (roads, rivers, city center, vegetation, wetlands, etc.). Algorithms used in this paper were

logistic regression and decision tree. Authors concluded that the logistic regression proved to be a better solution for their problem and believe that their work, as well as similar research in this field, can bring great benefit to society when it comes to fire prediction.

A group of authors from China presented a problem of fire prediction [3] in a region of China called Heilongjiang, where the main focus of their work was to determine which factors have the greatest influence on fires when determining the probability of forest fires and finally how to determine different fire prediction mechanisms. Several data sets were mentioned in this paper, with the most significant one being the one that contains detailed information about each forest fire, including exact geographic location, size of the fire and its location. What is interesting about this paper is that the authors focused only on logistic regression as they considered it being suitable algorithm to solve the given problem. What they focused more on was the analysis of the possible causes of the fire since they believed it would lead to a better result. This paper also deals with the climatic factors on fires and thanks to this, the authors were able to determine the causes of the fires in parts of the Heilongjiang region and the probability of fire outbreaks in a certain part of the region.

In the paper [4] group of authors from Slovenia highlighted their motivation for fire prediction in Slovenia and their engagement in various data-mining techniques. Although this paper is published a while ago, its importance is significant as it is one of the most relevant and cited papers in this field. Three sets of data were used in this research and as the authors point out, each one of them contained information about some area of Slovenia: Kras, coastal and continental area. Each of these data sets contained a lot of geographic data such as: area of land under forest, urban area land, etc. They also contained climate data such as: average temperature for a certain day, air humidity, amount of precipitation, wind speed and direction, etc. The authors point out that they had a lot of data and that after analyzing it, they approached the problem by using different algorithms. They concluded that the logistic regression was useful to find out dependencies between the parameters, but when it comes to the prediction itself best results were achieved using bagging algorithms where all algorithm parameters were predefined values (they were not optimized). The authors state that the reason for non-optimized use of parameters for various algorithms is the fact that the main goal of their work was to learn what could be used to solve the problem of fire prediction and not optimization itself.

Authors from China [5] used data from multiple sources and combined data about fire hotspots, meteorological data, terrain, vegetation and socio-economical data for the period from 2003 and 2016 for the purpose of forest fire prediction in China. By using an optimized model, they obtained best result by using random forest algorithm. On the other hand, authors of the paper [6] used AdaBoost classification algorithm with the idea to predict the occurrence of forest fires and they compared performance and results with SVM and Decision tree algorithm. They concluded that AdaBoost is a promising approach in predicting the size of the fire.

Since none of the mentioned authors tried to determine how much some of the regions in some period would be threatened (affected by the fires) we decided to focus on this angle in our research. Such solution could be extremely useful, as the society

and competent authorities would be aware of how much a certain region is at risk in each period of the year and thus enable the authorities to plan fire protection.

3 Data sets and preprocessing

3.1 Data sets

Both data sets used in this paper are published by the Brazilian ministries and are available for free use. The first data set contains information on how many fires each of the 27 federal states of Brazil had in each month from 1998 to 2017. The features of this data set are:

- Ano - year when the fire happened,
- Estado - name of the federal state where the fire happened,
- Mês - month for which the number of fires was recorded,
- Nùmero - recorded number of fires,
- Perìodo - date when the number of fires was published.

Second data set contains geographic and demographic information for each of Brazil's 27 federal states. This data set contains features:

- Flag - flag of the federal state,
- Federative unit - name of the federal state,
- Abbreviation - abbreviation for the name of the federal state,
- Capital - capital city name of the federal state,
- Area (km^2) - area of the federal state expressed in square kilometers,
- Area (sq mi) - area of the federal state expressed in square miles,
- Population - population of the federal state,
- Density (per km^2) - population density of the federal state expressed in the number of inhabitants per square kilometer,
- Density (per sq mi) - population density of the federal state expressed in the number of inhabitants per square mile,
- GDP - Gross Domestic Product of the federal state expressed in billions of Brazilian reals,
- GSP per capita - Gross Domestic Product per inhabitant of the federal state expressed in Brazilian reals per inhabitant,
- HDI - Human Development Index of the federal state,
- Literacy - literacy of residents of the federal state,
- Infant mortality - infant mortality rate in the federal state,
- Life expectancy - the average length of life in the federal state,
- Statehood - the year when the federal state received the status of the federal state.

3.2 Exploratory data analysis and selection of the important features

Before merging data sets, encoding needed to be changed since a lot of data was written using characters that are specific for Portuguese language. Sets were merged

based on a common feature represented by the federal state of Brazil. In this data set, which consists of many features, it was necessary to identify which features are mutually dependent and which significantly affect the number of fires in the regions. For the exploratory analysis of the data was used RapidMiner software [7] since it has a set of useful tools for data analysis. During the analysis process, the impact of various characteristics on the total number of fires was analyzed for the whole country of Brazil as well as for each federal state separately. Additionally, the trend in the number of fires was examined to find possible repetition in the data and to identify which of the existing data could influence the higher number of fires in some parts of the year.

When analyzing the data set, it was concluded that the number of fires does not highly depend on the year in which it was reported and that there is no trend in the data that would show repeating pattern every couple of years. On the other hand, data showed high correlation of the number of fires with the month in which the fire occurred. Data clearly shows that in some months the number of fires is significantly higher. It can be concluded that that the fire season starts in June and ends in November, with a peak in September when the highest number of fires is regularly recorded. The larger federal states in Brazil are those that contain larger areas and lower population density (mostly the Amazon Forest region), while the smaller ones are predominantly urban and with higher population density. Higher number of fires are recorder in greater states that have lager area, most of which is under the forests.

After analyzing the data set, a selection of significant features was made by using correlation coefficients [8]. Features that have been shown to have a higher correlation coefficient with the number of fires are:

- The month in which the number of fires was recorded,
- The area of the federal state,
- Density of the population,
- GDP per inhabitant of the federal state,
- HDI of the federal state,
- Literacy of residents of the federal region,
- Infant mortality rate in the federal region,
- Average life expectancy in the federal state.

Of the features listed above, only the first two along with the year when the number of fires was recorded were taken into further consideration. Other features were rejected because, although they have a high level of correlation, they did not help in making results more precise and included significant drop in performances for certain periods. Federal states that are larger in area and are mostly under the forests are less developed, less populated and have a weaker infrastructure. Due to less development, the HDI, population literacy and average life expectancy are lower than in more urban and developed federal states, while the infant mortality rate is higher. As these are characteristics of less developed federal states which also have the characteristic of having a greater number of fires due to geographical characteristics, it was shown that these characteristics are correlated with the recorded number of fires. However, since

these characteristics have no essential connection with the occurrence of fire, they were neglected in the final stages of research.

4 Methodologies

Methodologies used for the classification problem of the federal states of Brazil according to the level of danger from forest fires are:

1. Ridge,
2. SVM (Support-vector machine),
3. Random forest,
4. Ada boost,
5. Extra trees,
6. Gradient boosting,
7. KNN (k-nearest neighbors),
8. Decision tree.

Some of the applied algorithms are basis or part of other used algorithms (within the ensemble) so they do not represent a new approach to the problem and the results cannot be expected to be significantly different or better. The idea was to test how those algorithms that are part of some ensembles work independently and to see if they would give better results at least on some part of the data set that can later be used for tuning the algorithms that use them within the ensemble.

Federal states of Brazil were classified according to the level of vulnerability into four classes that represent different levels of vulnerability. Classes were empirically defined after analyzing the data. Based on the number of fires, following classes were defined:

- Class 1 - no danger (extremely small number of fires),
- Class 2 - low risk (moderately small number of fires),
- Class 3 - high risk (large number of fires),
- Class 4 - extreme high risk (extremely large number of fires).

The range of classes were defined after studying the data where we first observed Brazil in whole and later every state independently. For entire Brazil, defined range of number of forest fires, for each class is:

- Class 1 - from 0 to 8 forest fires,
- Class 2 - from 9 to 51 forest fires,
- Class 3 - from 52 to 248 forest fires,
- Class 4 - more than 248 forest fires.

For each of federal states separately classes were defined, and they differ between the states. For example, Mato Grosso is one of the largest states, covered with forests and therefore severely threatened by the fires. This state is classified with the Class 4, when number of fires in a month reaches over 1402 fires. On the other hand, federal state Sergipe has significantly lower number of forest fires and is assigned to the Class 4 when the number of fires is above 17. For each of the algorithms, many hy-

perparameters were tested and adjusted to give the best possible solution after cross-validation which was conducted using k-fold[1] principle.

5 Results and discussion

All previously mentioned algorithms were applied to the prepared data set and some that gave the best results for the given problem were singled out and will be discussed in more detail afterwards. Implementation of all the algorithms mentioned in the previous section is from scikit-learn [9] library. Accuracy was used to evaluate obtained results since it is often used in this field.

Calculating the accuracy of the prediction to which class the federal state belongs in a certain period was performed using accuracy metric and the results are shown below in the Fig 1. The average measurement values are listed for each algorithm in the table below. As the training and the test set are divided randomly, due to the different selection of the sets, different results were obtained for each prediction, and therefore the average values of those predictions were calculated.

Random forest algorithm was the one which performed the best as it was the case with some papers in this area. Similar accuracy was given by several other algorithms as listed in the table. Much was expected from the SVM algorithm due to it being praised in the relevant literature, but it did not live up to the expectations. KNN surprisingly achieved a high result, and this was not excepted because not that many papers included it as an algorithm with relevant results.

[1] k-fold - cross-validation statistical method used for estimation of skill of the machine learning models.

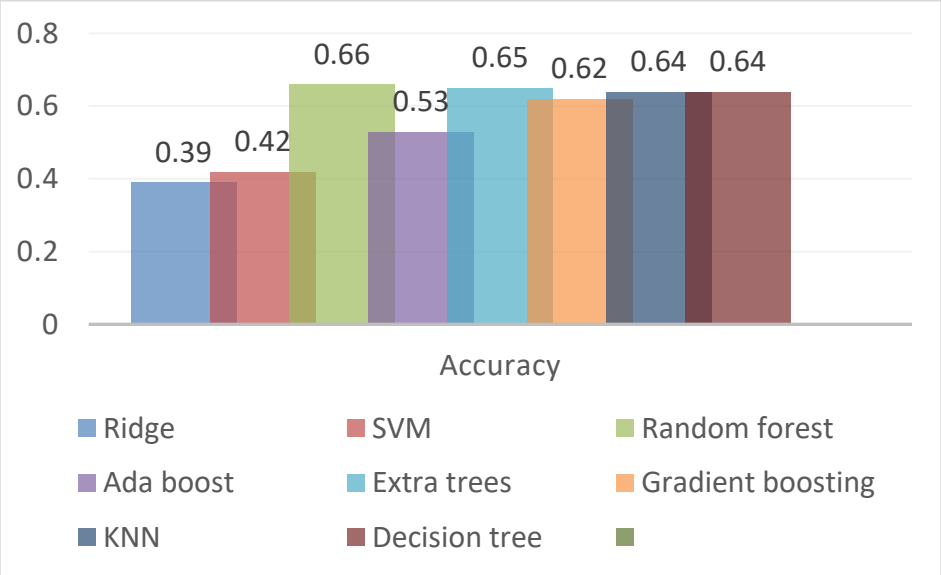

Figure 1. Results of classification using different algorithms.

6 Conclusion

This paper describes the problem of classification with the goal to classify federal states of Brazil by vulnerability and by that determining the degree of danger of these states from the forest fires. This kind of classification could help competent institutions and authorities that fight against disasters of this type every day.

Based on the data set on the number of forest fires for each state and each month from 1998 to 2017 and data set of geographical and demographical data on federal state that were joined based on common feature (federal state), an exploratory analysis was performed, and feature correlation was studied. Furthermore, several different models were trained, and the results of these models were analyzed. A certain group of models based on mostly similar algorithms gave similar results in the end. Some of the often-used models in this field, like SVM, did not give expected results. Best results were obtained with random forest algorithm according to the accuracy metric.

Further research in this area could be directed towards the collection and incorporation of data on meteorological conditions and their impact on the occurrence of forest fires. Many papers in the field of forest fires, especially those investigating the most significant causes of forest fires and their prediction, base their research on meteorological data. Although we wanted to include this kind data into our research, unfortunately meteorological data was not regularly and precisely collected for all federal states of Brazil since 1998 with most data missing for less developed states.

Only the data for the last few years is publicly available and that could be used in the future for the purposes of solving the earlier mentioned problem.

Another direction of the future research could be the study of anthropomorphic (human) influence on the cause of forest fires. With the increase in fires of this type, there is notable increase in number of published papers in the field of forest fire prediction that are trying to find correlation between the number of forest fires and specific human activities that could likely cause them. Unfortunately, due to insufficient research in this area and lack of data, there have been no major success so far which will hopefully change in the future.

References

1. Shidik, Guruh Fajar, et al. "Linked open government data as background knowledge in predicting forest fire." *Jurnal Informatika* (2014).
2. Sitanggang, I. S., et al. "Predictive models for hotspots occurrence using decision tree algorithms and logistic regression." *Journal of applied sciences* 13.2 (2013): 252-261.
3. Chang, Yu & Bu, Rencang & Chen, Hongwei & Feng, Yuting & Li, Yuehui & Hu, Yuanman & Wang, Zhicheng. (2013). Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. Landscape Ecology. 28. 10.1007/s10980-013-9935-4.
4. Stojanova, Daniela & Panov, Pance & Kobler, Andrej & Džeroski, Sašo & Taškova, Katerina. (2006). Learning to predict forest fires with different data mining techniques.
5. Pang, Yongqi, et al. "Forest fire occurrence prediction in China based on machine learning methods." Remote Sensing 14.21 (2022): 5546.
6. Rosadi, D., and W. Andriyani. "Prediction of forest fire using ensemble method." Journal of Physics: Conference Series. Vol. 1918. No. 4. IOP Publishing, 2021.
7. RapidMiner, https://rapidminer.com/, last accessed 2023/12/25.
8. Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer, Berlin, Heidelberg.
9. Scikit-learn library, https://scikit-learn.org/stable/, last accessed 2024/01/03.