Instituto Superior de Agronomia ULisboa

Master's in Green Data Science 2023-2024

Practical Machine Learning/Aprendizagem Automática Aplicada

Instructor: Manuel Campagnolo

Students: Alícia Eva Lopes Gouveia (27900) & Sofia Margarida Matias Rodrigues (27899)

Introduction

Portugal is inserted into the Mediterranean region where the climatic regime of mild, wet winters and warm, dry summers influences the recurrent fires. Global warming is reshaping Europe's pyroregions and climate change-driven fires can differ from natural fire regimes. Extreme fire seasons are becoming longer, more common and increasingly unpredictable [2].

Forest fires are a worldwide occurrence and have major effects on the ecosystem and environment [3]. In recent years, the world forest disturbance regimes have intensified, and future climatic changes are expected to amplify this development further in the coming decades [4]. A lot of factors consequent to climate variability contribute to raise risk and influence forest fire behaviour such as heat waves, changing regional weather patterns, droughts and temperature fluctuations [3].

Forest fires play a major role in Portugal's environment, their impact leads to the destruction of forest and other negative effects surrounding biodiversity. Forests provide a vast of direct and indirect benefits to humankind therefore, disasters like wildfires have a major impact. These wildfires are uncontrolled fires that spread rapidly through vegetation, affecting forests, grasslands, and agricultural areas. They are a significant environmental hazard that can lead to loss of life, property damage, and ecological destruction.

In this *Forecast Forest Fires* project, we aim to develop a method for early fire detection that can help decision makers plan mitigation and extinguishing tactics and help authorities and emergency services to be alert and prepare accordingly.

Understanding wildfires through detailed analysis of historical data is crucial for improving fire management strategies. By leveraging our dataset, we can gain insights into the factors driving fire occurrences, the effectiveness of interventions, and the impact of various conditions on fire behaviour. This knowledge can inform policies and practices to enhance fire prevention, detection, response, and recovery efforts in Portugal, ultimately reducing the adverse impacts of wildfires on communities and the environment.

Data

In this project we will work with a Portuguese dataset provided by ICNF (Instituto da Conservação da Natureza e das Florestas) with information regarding fire occurrences in Portugal, throughout the years of 2011 and 2020. It was obtained here: ICNF - Instituto da Conservação da Natureza e das Florestas.

The dataset is a .xlsx file composed by 41 columns and 177 129 entries. These columns can be divided between temporal data, geographical data, fire characteristics and meteorological indices. All of those entries are associated with a fire occurrence identified by the table's primary key - Codigo_SGIF. The dataset attributes and their definitions are described as seen on Appendices.

For the purpose of this machine learning problem, we decided to choose certain variables from the dataset, those being: DSR, FWI, ISI, DC, DMC, FFMC, BUI.

DSR is calculated based on FWI (Canadian Forest Fire Weather Index System) and these values can range from 0 to and above 64; ISI has values that range from 0 to 30; DC we have values in the dataset that range from 0 to 1043; DMC the values range from 0 to 467; FFMC values range from 0 to 99 and BUI values range from 0 to 325 (the dataset has values up to 457). [1]

We opted to implement a Classification approach with this model, where our goal is to obtain an application that will output Fire or No Fire options after attributing the values for each of the selected variables.

Therefore, we start our code by creating a binary target variable using the "AreaTotal_ha" column which represents the total burned area. If this is greater than 0 it will indicate Fire and otherwise it indicates No fire, and creates a new column called "Fire".

To reduce the computational load, we implemented a step for sampling the data. Here, a smaller fraction of the dataset (10%) was used in order to reduce the computational resources needed for data processing, model training and evaluation. This allows for a faster execution time while performing Exploratory Data Analysis (EDA) and initial model development.

It is also important to handle missing values so a step was added to the code where any missing values in the dataset are dealt with and ensure that the subsequent data processing and model training step can proceed without errors. Missing values can cause issues with many machine learning algorithms and data analysis techniques, which generally require complete data. To perform an EDA, handling the missing values helps generate accurate plots and statistics and ensures the dataset is suitable for training models.

The Exploratory Data Analysis is implemented to plot histograms of the numerical features (meteorological indices) to understand their distribution, detect outliers and anomalies, uncover patterns and gain insight that can guide the choice of algorithms.

We also obtained a Correlation Matrix to visualize the relationships between features and the target variable. High positive or negative correlations (1 or –1) indicate strong relationships between features. Identifying features with high correlation to the target variable helps indicate good candidates for predictive modeling and this high correlations between features can indicate multicollinearity, which might require dimensionality reduction techniques like PCA (Principal Component Analysis).

We also addressed the class imbalance in the dataset. This imbalance occurs when one class significantly outnumbers the other(s) which can lead to biased models that perform poorly. For this purpose, we used the technique Synthetic Minority Over-sampling Technique (SMOTE) which works by creating new instances that are combinations of existing minority class instances. This balances the dataset, prevents oversampling (SMOTE creates new, unique samples) and then improves the model performance.

Data Organization

Splitting the dataset into training and testing sets is a crucial step in machine learning for it allows to train the model (using the training set) and to evaluate the model (using the testing set to evaluate the model's performance on unseen data). The training set allows the model to learn from part of the data and the testing set allows to evaluate the model's performance on data it has not seen before, providing an unbiased assessment of its generalization ability. We opted for using 70% of the data as a training set and 30% of the data as a testing set. By splitting the data, we ensure that our machine learning pipeline can efficiently train the models as it provides a robust measure of how well the models will perform in real-world scenarios.

Before splitting, the dataset has been pre-processed, imputed for missing values and balanced using SMOTE (as written above).

Methods

For this project, two machine learning models are used: Logistic Regression and Random Forest Classifier. They serve as classification models but differ in terms of their approach, and performance characteristics. Both models are trained on PCA - transformed data to reduce dimensionality and enhance computational efficiency, and their performance is evaluated using standard classification metrics.

The Logistic Regression is a linear algorithm used for binary classification tasks and it models the probability that a given input belongs to a particular class using a logistic function (sigmoid). For this model the default hyperparameters were chosen to prevent overfitting. A random parameter was implemented to ensure that the random processes involved in training the model (such as shuffling the data) produce the same results each time the code is run. This is especially important for debugging and comparing model

performance across different runs. Other techniques implemented were regularization techniques where a penalty was added to the loss function.

The Random Forest Classifier combines multiple decision trees during training and outputs the classes for classification to improve the performance and robustness.

The Random Forest is likely to perform better due to its ability to handle complex patterns and interactions in the data. However, Logistic Regression provides valuable insights into feature importance and relationships, making it a useful baseline model.

By implementing both models, we can compare their performance using various evaluation metrics. This helps understand which model is better suited for the specific dataset and problem. This dual implementation aids in selecting the best model for predicting fire occurrences in forests and provides a robust framework for future improvements and analyses.

Because we are trying to create a classification system for early fire detection, we opted to also create an application where it is possible to input the values for the meteorological indices and see if the result is Fire or No Fire. As explained before, this application can help authorities and other entities to predict fire occurrences and, in that way, manage the areas to prevent them to happen or to reduce the severity of the fires.

Results

The resulting plots, of the plot distribution, correlation matrix and PCA analysis were placed in the Appendices field.

For the evaluation of the performance of the two used models, some metrics were used, such as: Accuracy, Precision, Recall, F1 score and ROC AUC. The Accuracy represents the ratio of correctly predicted instances in the total, Precision is also a ratio, but used to indicate the correctly predicted positive observations to the total (predicted positives), the Recall is the ratio of correctly predicted positive observations to all observations in the actual class. The F1 score is the weighted average of precision and recall and ROC AUC is the area under the receiver operating characteristic curve, which is a measure of the model's ability to distinguish between classes.

The results that represent ratios could also be converted to percentages but are shown in the following table as they were presented in the output from the code.

Parameter	Logistic Regression	Random Forest
Accuracy	0.56498	0.92494
Precision	0.57414	0.94722
Recall	0.50067	0.89990
F1 Score	0.53489	0.92296
ROC AUC	0.56493	0.92492

Analysis

The results for the respective models show some important parameters to evaluate the model's performance. In this case, it is possible to say that the Random Forest model applied to the dataset performs well when compared to the Logistic Regression, due to the higher values.

The results from the EDA analysis were histograms that represent each of the chosen indices. Daily Severity Rating (DSR) indicates the potential difficulty in controlling a wildfire if one were to start. The graph for DSR shows that most values of DSR are low, with a high concentration around 0-10. This implies that on most days, the fire control difficulty is relatively low. The Fire Weather Index (FWI) indicates the fire intensity, and its graph indicates that most FWI values are between 10 and 40, suggesting moderate fire intensity under typical conditions. Initial Spread Index (ISI) measures the expected rate of fire spread immediately after ignition and its graph shows that most ISI values are between 0 and 10, indicating that initial fire spread is generally slow to moderate. The Drought Code (DC) indicates long-term moisture content of deep, compact organic layers and the corresponding graph shows that the values are spread across a wide range, with a concentration between 0 and 600. Higher DC values suggest a significant number of days with drought conditions affecting deeper soil moisture. For the Duff Moisture Code (DMC) it reflects the moisture content of compact organic layers, and most DMC values are clustered between 0 and 100, implying that the shallow organic layer typically have moderate moisture content. The Fine Fuel Moisture Code (FFMC) indicates the moisture content of fine surface fuels, and the graph demonstrates high concentration around 80-90 suggesting that fine surface fuels are often dry and thus easily ignitable. The Buildup Index (BUI) represents the total amount of available fuel for combustion and its graph shows that most BUI values are between 0 and 200, indicating moderate amount of combustible material available under typical conditions.

The histograms indicate the frequency distribution of these indices over the sampled dataset. Most indices show a right-skewed distribution, meaning that extreme values are less common but still possible.

The Correlation Matrix helps understanding how different indices interact and influence each other. We obtained high correlations between certain indices, like DSR and FWI, or DMC and BUI, indicate redundancy. The DSR is strongly correlated with FWI (0.97) which indicates that as the daily severity rating increases, the fire weather index also increases significantly. The FWI is strongly correlated with DSR (0.97) and ISI (0.87) which suggests that FWI is closely related to both the severity rating and the initial spread. The ISI shows lower correlations with DC (0.30) and DMC (0.43) which suggests less dependence on long-term moisture content. The DC (Drought Code) shows strong correlation with BUI (0.84) and it suggests that drought conditions significantly affect the amount of available combustible material. The DMC shows very strong correlation with BUI (0.98), an almost perfect correlation, indicating that the

moisture content of shallow organic layers is a major component of the available fuel index. The FFMC shows lower correlation with DMC (0.38) and BUI (0.39), and it suggests less influence from overall moisture content and available fuel. The BUI is strongly correlated with DC (0.84) and DMC (0.98) which highlights its dependence on the long-term moisture content of the organic layers.

Regarding the PCA, the plot indicates that the first two principal components are sufficient to capture the majority of the variance in the dataset. This ensures that the most significant information is retained while reducing the complexity. Four Principal Components were implemented where the first captures most of the significant patterns in the data (70%), the second around 20%, the third and fourth with less variance, around 7% and 3%, respectively. All added brings it around 100%

The performance of the application that we created encountered some problems, because it is not predicting the correct output when we input some values on the app. This problem may be occurring due to using the Scaler and PCA transformations, but it is necessary to have them, because the same pre-processing steps need to be applied. This ensures that the input features are in the same format that the model expects.

References

[1] "Instituto Português Do Mar E Da Atmosfera." *Www.ipma.pt*, www.ipma.pt/pt/riscoincendio/fwi/. Accessed 5 July 2024.

[2] Latham, Katherine. "Wildfires: The Changing Face of the Mediterranean Landscape." *Www.bbc.com*, www.bbc.com/future/article/20230803-wildfires-the-changing-face-of-the-mediterranean-landscape.

[3] Stavros Kalogiannidis, et al. "Socio-Psychological, Economic and Environmental Effects of Forest Fires." *Fire*, vol. 6, no. 7, 21 July 2023, pp. 280–280, https://doi.org/10.3390/fire6070280. Accessed 23 Aug. 2023.

[4] Thom, Dominik, and Rupert Seidl. "Natural Disturbance Impacts on Ecosystem Services and Biodiversity in Temperate and Boreal Forests." *Biological Reviews*, vol. 91, no. 3, 22 May 2015, pp. 760–781, www.ncbi.nlm.nih.gov/pmc/articles/PMC4898621/, https://doi.org/10.1111/brv.12193.

Contributions

All the processes necessary to implement this project were done by both members. The main project management system was trial and error, and both members tried to get different machine learning models to see which better apply to the problem in question.

Appendices

Attribute Description (note: this was taken directly from the dataset so it's in Portuguese)**:**

Codigo_SGIF- Identificador único SGIF do incêndio rural

Codigo_ANEPC- Identificador da ocorrência na base-de-dados SADO da ANEPC (igualmente designado por código NCCO)

Ano- Ano civil da data de alerta do incêndio

Mes- Mês da data de alerta do incêndio

Dia- Dia da data de alerta do incêndio

Hora- Hora do alerta do incêndio

AreaPov- Área ardida em povoamentos florestais (ha)

AreaMato- Área ardida em matos (ha)

AreaAgric- Área ardida em zonas agrícolas (ha)

AreaTotal- Área ardida total (povoamentos + matos + zonas agrícolas)

ClasseArea- Classe de área ardida total (ha)

DataHoraAlerta- Data/hora de alerta do incêndio

DataHora_PrimeiraIntervencao- Data/hora da primeira intervenção

DataHora_Extincao- Data/hora da extinção

Duracao_Horas- Duração do incêndio (período entre a data/hora de alerta e a data/hora de extinção)

IncSup24horas- Incêndio rural com duração superior a 24 horas (período entre a data/hora de alerta e a data/hora de extinção)

DTCCFR- Código INE da freguesia de início do incêndio

Distrito- Distrito do ponto de início do incêndio

Concelho- Concelho do ponto de início do incêndio

Freguesia- Freguesia do ponto de início do incêndio

Local- Local do ponto de início do incêndio

RNAP- Local de início situado numa área da Rede Nacional de Áreas Protegidas

RNMNPF- Local de início situado numa área da Rede Nacional de Matas Nacionais e Perímetros Florestais

X_Militar- Coordenada X do ponto de início do incêndio em Datum Lisboa Hayford Gauss (EPSG: 20790)

Y_Militar- Coordenada Y do ponto de início do incêndio em Datum Lisboa Hayford Gauss (EPSG: 20790)

Latitude- Latitude do ponto de início do incêndio (unidade: graus decimal)

Longitude- Longitude do ponto de início do incêndio (unidade: graus decimal)

X_ETRS89- Coordenada X do ponto de início do incêndio no sistema de coordenadas PT-TM06/ETRS89 (EPSG: 3763)

Y_ETRS89- Coordenada Y do ponto de início do incêndio no sistema de coordenadas PT-TM06/ETRS89 (EPSG: 3763)

DSR- Índice de perigo meteorológico de incêndio rural (calculado com base no FWI)

FWI- Índice meteorológico de perigo de incêndio rural

ISI- Índice meteorológico de propagação inicial do fogo

DC- Índice meteorológico de seca

DMC- Índice meteorológico de humidade da manta-morta

FFMC- Índice meteorológico de humidade do combustível fino

BUI- Índice meteorológico de combustível disponível

CodCausa- Código da causa do incêndio | ver correspondência na folha "TipoCausa"

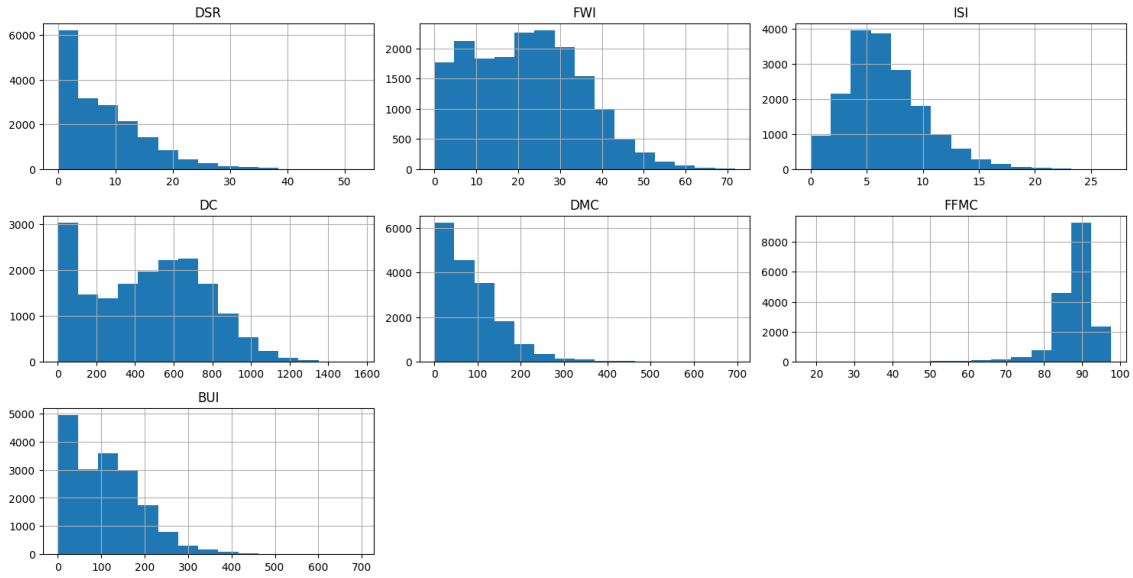
TipoCausa- Tipificação da causa

GrupoCausa- Grupo da causa

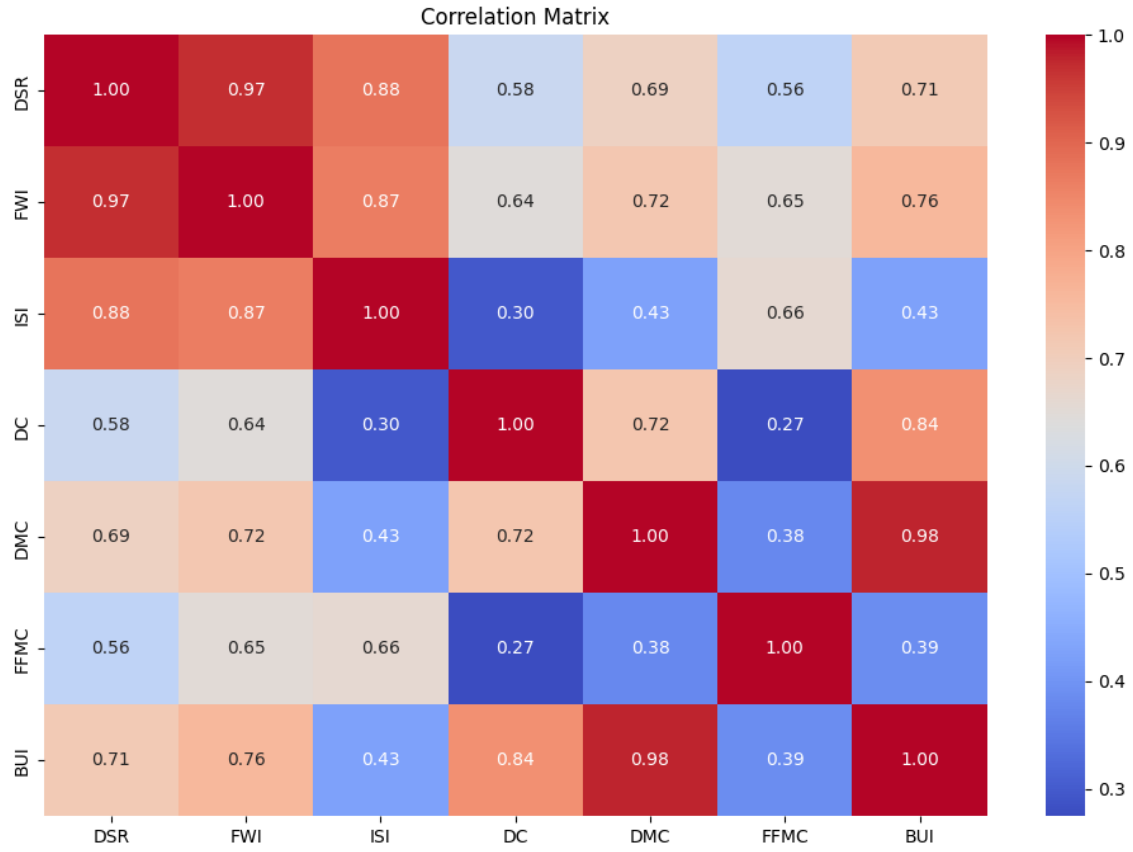
DescricaoCausa- Descrição da causa

Fonte de alerta - Primeira fonte a detetar/alertar

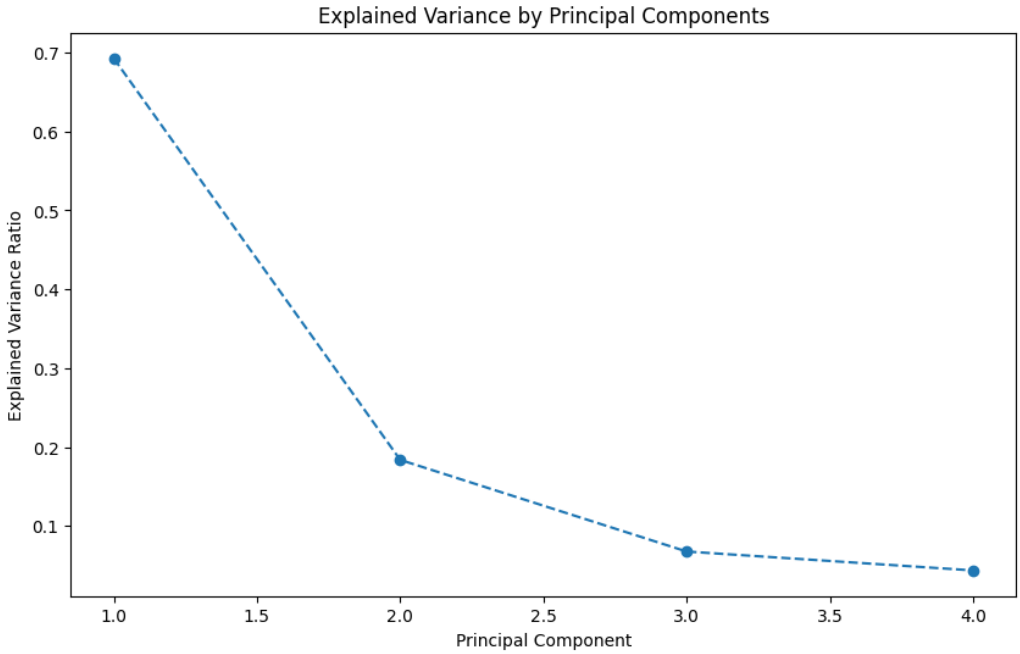
EDA Analysis:



Correlation Matrix:



PCA Analysis:


Explained Variance by Principal Components
Explained Variance Ratio
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component