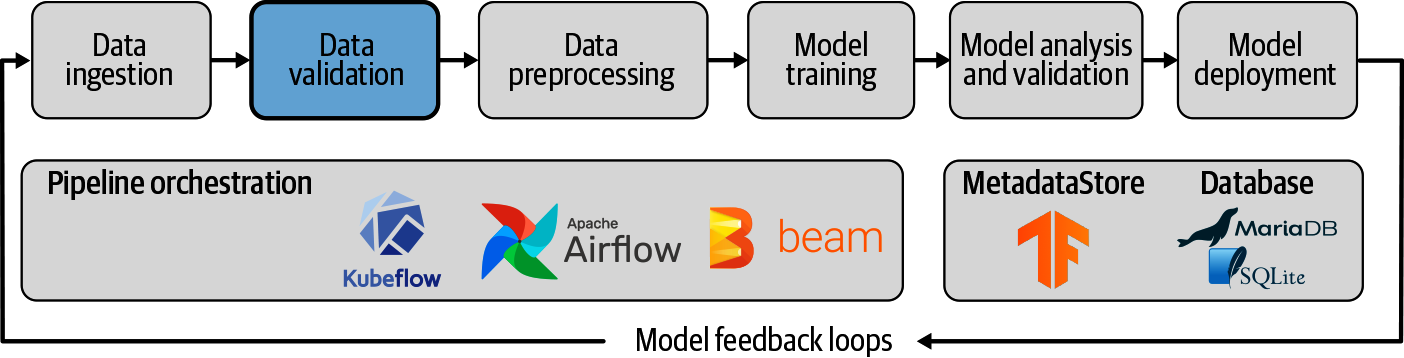
**Welcome to the Predictor for burned areas by forest fires**

**Overview**

The target is to predict burned areas using a regression model. Dataset file: data/forestfires.csv We'll use the Forest Fires Dataset (Indices from the Fire Weather Index System: Drought Code DC, Duff Moisture Code DMC and Fine Fuel Moisture Code FFMC)

by Christian Ramirez

**Tasks included:**

1. Analyze and understand the provided dataset through Exploratory Data Analysis.
   * Understanding the Variables:
     + X and Y: Spatial coordinates
     + month and day: Time variables indicating when the fire occurred
     + FFMC: Fine Fuel Moisture Code, a measure of the moisture content of litter and other fine fuels
     + DMC: Duff Moisture Code, a measure of the moisture content of loosely compacted organic layers
     + DC: Drought Code, a measure of the moisture content of deep, compact organic layers
     + ISI: Initial Spread Index, a measure of the expected rate of fire spread
     + temp: Temperature in Celsius degrees
     + RH: Relative Humidity (%)
     + wind: Wind speed (km/h)
     + rain: Precipitation in mm
     + area: The target variable, the burned area of the forest fire (in hectares)
   * Distributions and Patterns:
     + The target variable area has a right-skewed distribution, with many small fires and a few large fires, The graph confirms the right-skewed distribution of the target variable area, with a long tail indicating the presence of a few very large fires. Most fires have a burned area of less than 10 hectares, while a small number of fires exceed 100 hectares or even 1000 hectares. So, while most fires are relatively small, a few large fires contribute significantly to the total burned area.
     + The meteorological variables (FFMC, DMC, DC, ISI, temp, RH, wind, rain) show varying distributions, some skewed and some more normal.
     + The time variables (month, day) indicate that fires occur throughout the year, but certain months (e.g., August, September) seem to have more occurrences.
   * Relationships with Meteorological Variables:
     + The scatter plots show the relationships between area and various meteorological variables.
     + There appears to be a positive correlation between area and variables like FFMC, DMC, DC, ISI, and temp. As these variables increase, the burned area tends to increase as well.
     + The relationship between area and RH (Relative Humidity) appears to be negative, with higher humidity associated with smaller burned areas.
     + The relationship between area and wind seems more complex, with both low and high wind speeds potentially leading to larger fires.
   * Temporal Patterns:
     + There is a clear seasonal trend, with the highest burned areas occurring during the summer months (June-August), and the lowest during the winter months (December-February).
     + This aligns with the expectation that drier and warmer conditions during summer contribute to increased fire risk and severity.
     + Additionally, there may be variations within each month, indicating the influence of specific weather conditions on a given day.
   * Implications:
     + The skewed distribution of area suggests that most fires are relatively small, but a few large fires contribute significantly to the total burned area. Proper resource allocation and preventive measures should be in place to handle both small and large fires.
     + The meteorological variables provide insights into the conditions that could influence fire behavior and severity. Understanding their relationships with area can help in developing predictive models and decision support systems for fire management.
     + The temporal patterns observed in the time variables (month, day) can be useful for identifying high-risk periods and implementing targeted prevention and preparedness measures.
     + Correlations between variables can guide feature selection and engineering processes, as well as highlight potential confounding factors that need to be accounted for in predictive models.
2. Determine the insight you want to address with a machine learning model.
   * "Can we accurately predict the burned area of a forest fire (target variable: area) given the meteorological conditions and other relevant variables?"
3. Identify why an MLOps strategy is needed for this dataset. (Module 1)
   * A MLOps strategy would be beneficial for deploying and maintaining a machine learning model for predicting the burned area of forest fires in the northeast region of Portugal. Here are some key reasons why a MLOps strategy is necessary:
   * Data Management and Preprocessing:
     + The dataset contains various types of data (numerical, categorical, spatial, temporal) that require proper preprocessing and feature engineering.
     + A MLOps strategy can help streamline the data ingestion, validation, and preprocessing pipelines, ensuring consistent and reliable data processing for model training and inference.
   * Model Training and Evaluation:
     + Training machine learning models, especially complex ones like ensemble models or spatial-temporal models, can be computationally intensive and time-consuming.
     + MLOps practices, such as automated model training pipelines, distributed training, and efficient resource management, can greatly improve the efficiency and scalability of the model training process.
     + Robust evaluation metrics and techniques (e.g., cross-validation, holdout sets) are crucial for assessing model performance accurately, which can be facilitated by a MLOps framework.
   * Model Deployment and Serving:
     1. Once a satisfactory model is trained, it needs to be deployed and served in a production environment for real-time predictions or batch processing.
     2. MLOps strategies involve containerization, orchestration, and automated deployment pipelines, ensuring that the model is consistently and reliably served without manual intervention.
   * Monitoring and Maintenance:
     1. The performance of the deployed model needs to be continuously monitored, as the underlying data distributions or patterns may change over time (e.g., due to climate change, policy changes, or other factors).
     2. MLOps practices include monitoring model performance, data drift detection, and automated retraining/redeployment pipelines to ensure the model remains accurate and up-to-date.
   * Scalability and Reliability:
     1. As the demand for fire risk predictions grows or the complexity of the model increases, the system needs to be scalable and reliable to handle increased workloads and ensure high availability.
     2. MLOps strategies involve load balancing, auto-scaling, fault tolerance, and other techniques to ensure the system can handle varying demand while maintaining reliability.
   * Collaboration and Governance:
     1. Machine learning model development and deployment often involve cross-functional teams (data engineers, ML engineers, domain experts, decision-makers).
     2. MLOps practices promote collaboration, version control, documentation, and governance, ensuring that the entire lifecycle of the machine learning system is well-managed and transparent.
4. Design the pipeline architecture for this new machine learning initiative.
   1. 
   2. We are currently at the model evaluation stage. We’ve already completed data preprocessing using pandas, there was no need for any feature engineering. For the model selection the one being use is Random Forest Regressor (although this could change in the coming weeks) and scikit-learn was used for data splitting, training, prediction and evaluation.
5. Create a baseline model to address prediction tasks (classification, regression, etc.) related to the question. This model does not need high accuracy, recall, or F1 score; the goal is to create a quick model for iteration. (Module 4)
   1. This can be reviewed at the src folder on main.py
6. Set up the correct model structure with the aim of preparing it for deployment. (Módulo 6)
   1. This can be reviewed at the github repository itself.