College/University Name: CICCC - Cornerstone International Community College of Canada **Course**: Data Science **Instructor**: Austin Eghbal **Student Name**: Amir Lima Oliveira **Submission Date**: 2025-06-09

Machine Learning Project: Dryness Trend Prediction for Wildfire Avoidance

🧠 Machine Learning Project Template

This document outlines the machine learning project for predicting dryness trends to avoid wildfires, incorporating the Canadian Forest Fire Weather Index (FWI) system theory. It follows a concise structure, focusing on key aspects of the ML lifecycle.

What is the goal?

Goal: To develop a machine learning model that accurately predicts dry weather conditions across different geographical locations, specifically to aid in wildfire prevention and proactive resource management. This involves defining \'dry\' based on meteorological parameters relevant to fire danger and building a predictive model to forecast these conditions.

Problem Statement: Unpredicted dry spells significantly increase wildfire risk, leading to devastating environmental and economic consequences. This project addresses this by leveraging machine learning to forecast dry conditions, providing critical early warnings for wildfire mitigation.

Canadian Forest Fire Weather Index (FWI) System Context

The Canadian Forest Fire Weather Index (FWI) System is a national standard for assessing forest fire danger in Canada. It consists of six components that account for the effects of fuel moisture and weather conditions on fire behavior. Understanding these components provides crucial context for defining \'dry\' conditions in this project:

- **Fine Fuel Moisture Code (FFMC):** A numeric rating of the moisture content of litter and other cured fine fuels. It indicates the relative ease of ignition and flammability.

- **Duff Moisture Code (DMC):** A numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. It indicates fuel consumption in moderate duff layers.
- **Drought Code (DC):** A numeric rating of the average moisture content of deep, compact organic layers. It indicates seasonal drought effects on forest fuels and smoldering potential.
- **Initial Spread Index (ISI):** A numeric rating of the expected rate of fire spread, based on wind speed and FFMC.
- **Buildup Index (BUI):** A numeric rating of the total amount of fuel available for combustion, based on DMC and DC.
- **Fire Weather Index (FWI):** A numeric rating of fire intensity, based on ISI and BUI. It serves as a general index of fire danger.

This project\'s definition of \'dry\' conditions is inspired by the meteorological inputs used in the FWI system (temperature, relative humidity, wind speed, and precipitation), aiming to identify conditions that contribute to high FWI values and thus increased wildfire risk.

What data would be needed?

The primary dataset is `weather_classification_data.csv`. Key features for predicting dryness for wildfire avoidance, inspired by FWI system inputs, include:

- **Temperature:** Higher temperatures contribute to fuel drying.
- **Humidity:** Lower relative humidity indicates drier air and fuels.
- **Wind Speed:** High wind speeds can accelerate fire spread and fuel drying.
- **Precipitation (%):** Low or zero precipitation is a direct indicator of dry conditions and fuel moisture.
- **Cloud Cover:** Clear skies often correlate with higher temperatures and lower humidity.
- **Atmospheric Pressure:** Certain pressure systems are associated with stable, dry weather.
- **UV Index:** High UV index indicates strong solar radiation, contributing to drying.
- **Season:** Seasonal variations influence dryness and fire risk.
- **Visibility (km):** Clear visibility can indicate dry air.
- **Location:** Geographical context influences local weather and fire patterns.

Definition of \'Dry\' Conditions (inspired by FWI inputs): For this project, \'Dry\' conditions are defined by a combination of low precipitation, low humidity, and high temperature. Specifically, a day is considered \'Dry\' if: `Precipitation (%) < 10`,

`Humidity < 50`, and `Temperature > 25`. These thresholds are illustrative and can be refined based on domain-specific knowledge or further analysis of fire danger ratings.

Which learning problem should be used?

This project uses **Supervised Learning**, specifically **Binary Classification**. We aim to classify weather conditions into two categories: \'Dry\' (indicating high wildfire risk) and \'Not Dry\'. This approach is suitable because we can create a labeled dataset where each instance is explicitly marked as \'Dry\' or \'Not Dry\' based on the defined criteria derived from FWI-related meteorological parameters. The model will learn to map input weather features to these binary labels.

What aspects of the data do you consider important?

Effective data preparation is crucial. The following aspects are important:

- **Basic Inspection:** Understand data types, column names, and overall dataset dimensions.
- **Data Cleaning:** Handle duplicate rows and missing values. For this dataset, duplicates will be removed, and missing values will be assessed (none found in initial inspection).
- **Outlier Handling:** Outliers in numerical features will be addressed using the Interquartile Range (IQR) method, specifically by capping extreme values. This approach helps to mitigate the impact of extreme values without removing data points, which is often preferred in environmental data where outliers might represent real, albeit rare, events.
- **Exploratory Data Analysis (EDA):** Visualize distributions of individual features (histograms), relationships between features and the \'Dryness_Label_Wildfire\' (box plots for numerical, count plots for categorical), and correlations among numerical features (heatmap). This helps in understanding patterns related to dryness.
    - **Statistical Tests and Correlation Choices:**
        - **T-tests:** Used to compare the means of numerical features (Temperature, Humidity, Precipitation (%), Wind Speed) between the \'Dry\' and \'Not Dry\' groups. A significant p-value (typically < 0.05) indicates that the mean of a feature is statistically different between dry and non-dry conditions, highlighting its relevance to dryness for wildfire avoidance. This helps confirm if the FWI-related inputs (temperature, humidity, precipitation, wind speed) indeed show significant differences when conditions are dry.

        - **Chi-squared Tests:** Applied to categorical features (Season, Location, Cloud Cover) to assess their association with the \'Dryness_Label_Wildfire\'. A significant p-value suggests a statistical dependency between the categorical feature and the dryness label, indicating that certain seasons, locations, or cloud cover types are significantly associated with dry conditions, which is crucial for wildfire risk assessment.
        - **Correlation Heatmap:** Visualizes the Pearson correlation coefficients between all pairs of numerical features. This helps identify strong linear relationships between variables. For instance, a strong negative correlation between Temperature and Humidity would be expected in dry conditions, aligning with FWI principles. Feature Engineering:*
    - **Target Variable Creation:** Define the \'Dryness_Label_Wildfire\' based on the FWI-inspired thresholds for precipitation, humidity, and temperature.
    - **Categorical Encoding:** Convert categorical features (`Cloud Cover`, `Season`, `Location`) into numerical format using One-Hot Encoding.
    - **Numerical Scaling:** Standardize numerical features using `StandardScaler` to ensure they contribute equally to the model.
- **Data Splitting:** Divide the dataset into training and testing sets (e.g., 80/20 split) with stratification on the \'Dryness_Label_Wildfire\' to maintain class distribution.

What algorithm(s) would be suitable?

For this binary classification problem, **Random Forest** is a highly suitable algorithm.

Justification: Random Forest is an ensemble method that builds multiple decision trees and combines their predictions. It is robust to overfitting, can handle non-linear relationships, and provides good performance on tabular data. It also offers insights into feature importance, which can help understand which weather parameters are most influential in predicting dryness for wildfire avoidance.

How would model performance be measured?

Model performance will be measured using classification metrics, with a strong emphasis on those relevant to wildfire prevention:

- **Recall (Sensitivity):** This is the most critical metric. It measures the proportion of actual \'Dry\' conditions that the model correctly identifies. High recall is essential to minimize False Negatives (missing a dry period), as failing to predict a dry condition can have severe consequences for wildfire prevention.

- **F1 Score:** The harmonic mean of Precision and Recall. It provides a balanced measure, especially important given that \'Dry\' conditions might be a minority class. A high F1 score indicates a good balance between identifying dry conditions and avoiding false alarms.
- **ROC AUC (Receiver Operating Characteristic - Area Under the Curve):** Provides an aggregate measure of the model\'s ability to distinguish between \'Dry\' and \'Not Dry\' classes across all possible classification thresholds. A higher AUC indicates better overall discriminative power.

What metrics should be used?

Beyond the primary metrics, the following tools and visualizations will be used for a comprehensive evaluation:

- **Confusion Matrix:** To visualize True Positives, True Negatives, False Positives, and False Negatives, providing a clear breakdown of prediction outcomes.
- **Classification Report:** A summary report showing precision, recall, and F1-score for both \'Dry\' and \'Not Dry\' classes.
- **ROC Curve:** A graphical plot illustrating the trade-off between True Positive Rate and False Positive Rate at various thresholds, helping to select an optimal operating point.
- **Precision-Recall Curve:** Particularly useful for imbalanced datasets, it plots precision against recall for different thresholds, focusing on the performance of the positive class (\'Dry\').
- **Error Analysis:** Systematically examining misclassified instances (False Positives and False Negatives) to identify patterns and areas for model improvement.

How can the model be deployed and maintained?

Deployment:

The model can be deployed as an **API (Flask or FastAPI)**. This allows other applications (e.g., wildfire management systems, public warning systems) to programmatically request dryness predictions by sending weather data. This approach offers scalability and seamless integration into existing or new digital infrastructures.

Maintenance:

Ongoing maintenance is crucial for the model\'s effectiveness:

- **Scheduled Retraining with New Data:** Periodically retrain the model with the latest weather data to account for concept drift (changes in weather patterns over time) and maintain accuracy.
- **Model Monitoring and Performance Drift Detection:** Continuously monitor the deployed model\'s performance metrics (e.g., Recall, F1 Score) in real-time. Implement alerts for significant drops in performance, indicating a need for retraining or re-evaluation of the model or data pipeline. This includes monitoring for data drift (changes in input data characteristics) and concept drift (changes in the relationship between input and output).

Conclusion

This project successfully outlines a machine learning approach to predict dryness trends for wildfire avoidance, a critical step in proactive environmental management. By integrating principles from the Canadian Forest Fire Weather Index (FWI) system, we established a clear and relevant definition of \'Dry\' conditions, directly linking meteorological data to wildfire risk. The detailed data preparation pipeline, encompassing basic inspection, cleaning, feature engineering, outlier handling, and comprehensive Exploratory Data Analysis (EDA) with statistical tests, ensures the robustness and reliability of the dataset for subsequent modeling.

The iterative process of refining the project goal and problem statement, coupled with the emphasis on concise and actionable documentation, highlights a practical and results-oriented approach to data science. The selection of Random Forest as a suitable algorithm, along with a focus on critical evaluation metrics like Recall and F1 Score, underscores the project\'s commitment to minimizing false negatives in wildfire prediction, where missed warnings can have severe consequences. Furthermore, the theoretical discussion on model deployment and maintenance emphasizes the long-term vision for this project, aiming for real-world impact through continuous monitoring and adaptation.

Overall, this project demonstrates a strong foundation for developing a predictive model that can contribute significantly to wildfire prevention efforts, showcasing the power of data science in addressing pressing environmental challenges.