FRE521D Final Project: Predicting Wildfire from Satellite Data

Available: January 24, 2025
Deadline: February 10, 2025
Presentation Date: February 12, 2025

Overview

This project is designed to provide you with a comprehensive understanding of building an end-to-end data science project. You will use Python for programming, SQL for data extraction, and Tableau for visualization, combining these tools into a complete data pipeline. By engaging in this project, you will gain hands-on experience in data preparation, machine learning modeling, and creating visualizations.

Project Objectives

The objective of this project is to predict wildfire occurrences using satellite data. Through this project, you will:

- Analyze and preprocess a real-world dataset to make it suitable for machine learning.
- Implement machine learning models to predict wildfire occurrences and evaluate their performance.
- Build a complete data pipeline using Python and SQL to automate and visualize the workflow.

Project Components

Part 1: Research Paper Analysis (5%)

- Read the following research paper: Predicting Wildfires from Satellite Images using Deep Learning.
- Deliverable: Submit a **summary (1-2 pages)** addressing the following:
    1. What problem is the paper trying to solve?
    2. What dataset and machine learning techniques were used?

3. What were the key results and findings?
4. Suggest one possible extension or improvement to the methodology.

Part 2: Identifying and Selecting Models (5%)

- Review and identify potential machine learning models for wildfire prediction. These could include:
  - Models covered in Weeks 1–3 of the course (e.g., regression, decision trees, or SVM).
  - Advanced models like MobileNetV3, ResNet50, or others discussed in external resources.
- Justify the choice of your model(s) and how they align with the project objective.

Part 3: Data Extraction and Preparation (5%)

- Extract the wildfire dataset using the following code:

```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("abdelghaniaaba/wildfire-prediction-dataset")

print("Path to dataset files:", path)
```

- Tasks:
  1. Clean the dataset, handle missing values, and prepare it for modeling.
  2. Perform exploratory data analysis (EDA) to identify trends and visualize key features.
  3. Engineer features if necessary to improve model performance.

Part 4: Machine Learning Modeling (10%)

- Implement the following steps:
  1. **Exploratory Data Analysis (EDA)**: Visualize data distributions, identify trends, and understand key features.
  2. **Data Preprocessing**: Normalize, scale, or transform the data, and split it into training and testing sets.
  3. **Model Training**:
     - Train your selected model(s) on the dataset.
     - Evaluate the model's performance using metrics such as accuracy, F1-score, and precision-recall.
     - Discuss your model's performance and suggest possible improvements.

Part 5: Data Pipeline and Visualization (5%)

- Build a pipeline that integrates:
  1. **Python**: Use Python for data preprocessing, modeling, and analysis.
  2. **SQL**: Store the processed dataset in an SQL database for better accessibility.

Part 6: Presentation (5%)

- On February 12, 2025, present your project to the class. The presentation should include:
  1. A summary of the machine learning model(s) you implemented and their performance.
  2. An overview of the data pipeline.
  3. Key findings and insights gained from the project.

Final Deliverables

- **Research Paper Summary (5%)**:
  - Summarize the paper in 1-2 pages, covering its goals, methods, findings, and a potential improvement or extension.
- **Project Report, Code, and Presentation (30%)**:
  - A comprehensive submission including:
    * Dataset extraction, cleaning, and preprocessing steps.
    * Details of the machine learning model(s) implemented, with results and evaluations.
    * The pipeline was built using SQL for data handling and visualization.
  - Include all Python scripts and SQL queries in your submission.
  - The group presentation will be assessed as part of the total marks.

Marking Rubric

Component	Weight
Research Paper Summary	5%
Data Extraction and Preparation	5%
Model Identification and Justification	5%
Machine Learning Modeling	10%
Data Pipeline (SQL)	5%
Presentation	5%
Total	**35%**

Submission Guidelines

- All deliverables must be submitted by **February 10, 2025 (11:59 PM)**.
- Include proper references for any publicly available resources used.
- Marks will be equally applied to all group members.