California Wildfire Prediction

Mariah N Cornelio

Project Proposal

- **Abstract:** This project predicts whether a wildfire will occur on a given day in each California region and additionally, what the weather looks like days prior to a fire, using historical weather and environmental data from 1984 to present day.
- **Introduction:** California experiences frequent and severe wildfires that threaten communities, ecosystems, and public health. These fires are driven by a combination of high temperatures, dry vegetation, low humidity, and strong winds. Understanding the environmental conditions that lead to wildfire ignition is critical for prevention and emergency response.
- **Motivation:** Climate change is accelerating. Many lives and homes have been taken from families. The 2020 fire season alone burned over 4 million acres. This project can be done because California has decades of high-resolution weather data available.
    - Yuan, X., & Wu, X. (2022). Machine learning-based wildfire risk mapping in California using remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*.
    - Abatzoglou, J. T., & Williams, A. P. (2018). Impact of anthropogenic climate change on wildfire across western US forests. *PNAS*.

Details

Though not a new problem, few have applied deep learning to it. I want to expand on existing algorithms using LSTM/GRU to see how weather conditions are days prior to an event, aiming to improve early warnings such as "Possible fire in two weeks". Previous works have used Decision Trees and Remote Sensing and other traditional ML models like LR.

- **GOAL OF PACKAGE:** Create an algorithm that predicts when a fire will occur in a California region and provide analytics of weather patterns days prior to an event using LSTM.
- **DATA**
  - Source: https://zenodo.org/records/14712845
  - Size: 14,989 rows and 14 features
  - Pre-processing: Needed
  - Target: FIRE_START_DAY (binary)
- **ML ALGORITHM**
  - Tools: Random Forest, XGBoost, LSTM, GRU, Temporal CNN
  - Packages: Scikit-learn, xgboost, lightgbm, catboost, tensorflow/keras, pandas, matplotlib, numpy
- **METRICS**
  - Baseline: Logistic Regression & Decision Tree
  - Evaluation: Recall (most important), precision, F1, ROC-AUC
  - Cross-validation will be done
- **WORKPLAN**
  - EDA → Preprocess (encode, scale, etc.) → Feature Engineering → Baseline model/Evaluate → LSTM model → Results/Interpretation