CSI 5130

Final Project Report

Predicting Forest Fires with K-Means Clustering & Random Forest Classifier

Harsirat Grewal

Abstract

Predicting forest fire risks is a critical task to mitigate damage, protect ecosystems, and inform emergency response strategies. This project explores the use of K-Means Clustering and Random Forest Classifier to forecast wildfire-prone regions. We replicated and extended the methodology from a prior study, applying it to a new weather dataset comprising 96,453 entries of hourly atmospheric data.

Introduction

In this project, we aim to forecast forest fire risks using weather data by leveraging K-Means clustering for label generation and Random Forest for classification. Our work builds on a methodology tested in Indonesia and adapts it to a broader dataset of global weather history. The goal is to evaluate how well this approach generalizes to different data sources and locations. A preprocessing pipeline was developed to clean and transform the dataset, removing missing values and standardizing formats. The data was clustered into five wildfire risk levels using K-Means, and the resulting clusters were used as labels to train a Random Forest Classifier. The final model achieved a perfect score of 100% on standard classification metrics. These results demonstrate the potential effectiveness of the two-stage unsupervised-supervised learning pipeline for early fire risk prediction. Future work will involve spatial integration using GIS and expansion to longer-term time-series forecasting.

Related Work

Multiple studies have explored machine learning for wildfire forecasting. Wood (2021) used data mining for fire area prediction with optimized matching, while Singh et al. (2021) developed a Parallel SVM for similar tasks. Elshewey et al. (2020) tested regression models on fire data.

Our method replicates the K-Means to Random Forest pipeline but tests its effectiveness on a different dataset and with refined preprocessing steps, showing the pipeline's robustness.

Data

The dataset used consists of 96,453 weather records from a file named '*weatherHistory.csv*'. Each entry represents an hourly observation and includes:

- Date and time
- Temperature
- Apparent temperature
- Humidity
- Wind speed and bearing
- Visibility
- Pressure
- Weather summary (categorical)

The data required preprocessing:

- Removed rows with missing data.
- Normalized numerical fields.
- Aggregated and resampled data for temporal consistency when needed.

Methods

<u>K-Means Clustering</u>

Applied K-Means to assign a fire risk label based on weather parameters:

- Input features: Temperature, Humidity, Wind Speed, Wind Direction, and Pressure.
- Number of clusters: 5 (Very Low to Very High Risk).
- Distance metric: Euclidean.
- Output: Each data point was labeled with a cluster indicating fire risk.

<u>Random Forest Classification</u>

Using the risk labels from K-Means:

- Split the dataset into 70% training and 30% testing.
- Trained a Random Forest Classifier using the same input features as above.
- Performed classification with majority voting from decision trees.

Experiments

Conducted the following experiments:

Clustering Quality

- Used Silhouette Score to validate intra-cluster cohesion.
- Visually inspected clusters via 2D projections (e.g., Temperature vs Wind Speed).

Classifier Evaluation

- Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
- Cross-validation performed to ensure generalizability.

Achieved:

- Accuracy: 100%
- Precision, Recall, F1: All 1.00 across classes.

Conclusion

Our implementation confirms that combining unsupervised clustering (K-Means) with supervised classification (Random Forest) is a powerful approach to wildfire risk prediction. The system reached perfect scores under current settings, but such outcomes should be validated with longer and more diverse data. Future work includes:

- Extending the dataset over multiple years.
- Integrating geospatial data using GIS.
- Deploying a real-time fire risk alert system.

References

Wood, D. A. (2021). Prediction and Data Mining of Burned Areas of Forest Fires: Optimized Data Matching and Mining Algorithm Provides Valuable Insight. Artificial Intelligence in Agriculture, 5, 24–42.

Singh, K. R., Neethu, K. P., Madhurekaa, K., Mohan, A. H. P. (2021). Parallel SVM Model for Forest Fire Prediction. Soft Computing Letters, 3, 100014.

Elshewey, A. M., Esonbaty, A. A. (2020). Forest Fires Detection Using Machine Learning Techniques. Journal of Xi'an University of Architecture & Technology, Vol. XII, Issue IX.