Early prediction of wildfire using Machine learning Algorithms

“Forests are the lungs of our land, purifying the air and giving fresh strength to our people.” ~ Franklin D. Roosevelt

Abstract:

What is the achievable pinnacle of accuracy in wildfire prediction through the application of AI? It's widely recognized that machine learning algorithms can be harnessed to train predictive models for wildfires. These machine learning algorithms are more scalable and accurate, less expensive, and automatable — all large advantages over traditional prediction techniques (for example, trained individuals manning fire lookout towers).

We trained a variety of machine and deep learning models using two data sets from distinct locations (Algeria and Portugal) covering a variety of different meteorological inputs such as temperature, humidity and fire weather indices. In the end, we showed that a deep learning convolutional neural network model performed best achieving 98% accuracy, while the baseline machine learning models were unable to surpass 60%.

Through an analysis of relevant literature, it can be concluded that the only way to achieve the most realistic results is to collect real time data from the local region. This can be obtained from several means including weather feed, low orbit satellite, and drones.

Introduction

Contemporary methods for detecting wildfires demand extensive human effort [7]. In certain regions, infrared cameras have been strategically positioned to identify wildfires from advantageous viewpoints. One notable initiative, ALERTCalifornia, encompasses a network of over 1,000 high-definition, pan-tilt-zoom cameras distributed throughout California [16]. These cameras operate around the clock, equipped with near-infrared night vision capabilities, enabling the monitoring of natural calamities like active wildfires. However, the reliance on human scrutiny to meticulously analyze camera feeds for fire detection comes at a considerable cost.

Satellite imagery constitutes another approach employed to identify smoke plumes emanating from fires that rise beyond the forest canopy. Nevertheless, the effectiveness of optical solutions can be impeded by factors such as cloud cover or dense tree cover. Furthermore, the process of detecting wildfires using satellite imagery can be time-consuming, often requiring several hours to provide actionable insights [11].

The integration of IoT, Internet of Things, devices presents an alternative mechanism for wildfire detection. The company Dryad Networks proposes a similar solution [12]. These devices can be affixed to trees within wilderness areas and leverage wireless networks to transmit data. This data stream could then be harnessed to identify the presence of wildfires. However, a significant obstacle arises in the form of limited internet connectivity prevalent in remote locations—precisely where many wildfires originate.Relying on sensors completely is also not cost effective. If fire is spread in the wilderness, all of the sensors will be burnt with the fire itself.

A potential solution to these challenges lies in the utilization of artificial intelligence, especially machine learning and deep learning. In order to detect the conditions that lead to smoke with sufficient lead time to safeguard lives and preserve the integrity of wilderness ecosystems, we need to have a system that can predict with higher accuracy. The application of ML algorithms, especially the K nearest neighbor, decision tree and random forest classifier, could enable the identification of subtle patterns and indicators that herald the onset of wildfires. This proactive approach has the potential to transform wildfire detection from a reactive and labor-intensive process into a proactive and timely endeavor.

By employing AI, we can harness its computational power to analyze diverse data sources from Kaggle, a repository of public datasets. Specifically, two datasets are analyzed:

forest fire data from Portugal and Algeria. The patterns in these dataset, when amalgamated and scrutinized by AI algorithms, may yield predictive models that can anticipate the emergence of wildfires with enhanced accuracy and efficiency. The ability to forecast wildfires in their nascent stages could lead to more effective resource allocation, early evacuation plans, and targeted firefighting efforts.

Literature Review:

Similar work has been done in an earlier paper, *"Using the Canadian Fire Weather Index (FWI) in the Natural Park of Montesinho, NE Portugal: calibration and application to fire management"*, 2002. The data in this paper appears to have been analyzed primarily from a fire management and fire danger assessment perspective. The authors are interested in calibrating the Canadian Fire Weather Index (FWI) system to assess fire danger and establish preparedness levels within the Natural Park of Montesinho in Portugal.
The key focus areas of the data analysis from this perspective include:

1. Fire Danger Assessment: The authors aim to assess the potential for wildfires in the Natural Park of Montesinho based on meteorological and fuel moisture conditions, as reflected in the FWI components. They are attempting to classify days into different fire danger categories to provide actionable information for fire management.
2. Threshold Definition: The paper discusses the selection of specific thresholds within the FWI system to categorize fire danger. These thresholds are essential for defining when certain fire management actions should be taken.
3. Operational Preparedness: The authors are interested in providing guidance for fire management activities based on the calibrated FWI system. This includes defining different preparedness levels and corresponding actions to be taken by fire management personnel.
4. Ecological Considerations: The paper mentions the ecological role of fire in the shrublands within the study area. Therefore, part of the analysis may consider how fire management decisions impact the ecosystem and ecological objectives.

5. Local Factors: The authors acknowledge the influence of local factors, such as topography, land use, and social demographics, which can affect fire behavior and management. The analysis may have considered how these factors interact with the FWI system.

The data analysis in this paper is conducted with the primary goal of improving fire management strategies within the Natural Park of Montesinho by calibrating the FWI system and providing actionable insights based on fire danger assessments.

Based on the provided text, here are some conceptual points and potential areas of improvements:

- Data Set Limitations: The paper acknowledges that the data set has a significant imbalance, with only 14% of days having fire activity. This imbalance could affect the validity of statistical analyses, and it's important to consider how this might impact the conclusions drawn from the research.
- Threshold Selection: The paper defines thresholds for fire danger classes based on percentiles of the FWI index. While this approach is common, it can be somewhat arbitrary and should be carefully justified. Additionally, the paper mentions that a cluster analysis suggested only four classes, yet five were defined. This inconsistency should be addressed.
- Applicability to Different Regions: The study focuses on the Natural Park of Montesinho in Portugal. While the results may be relevant for this specific area, it's important to discuss the potential limitations of applying these findings to other regions with different climate, vegetation, and fire regimes.
- Influence of Local Factors: The paper mentions that local factors, including land use, landscape patterns, and social demographics, play a role in the fire regime. It's essential to elaborate on how these factors were considered or incorporated into the analysis and calibration of the FWI system.
- Validation: While the paper describes the calibration process, it's crucial to address how well the calibrated FWI system actually performs in predicting fire danger or fire occurrence in subsequent years or seasons. Validation of the model's accuracy is essential.

- Model Complexity: The paper briefly mentions logistic regression's failure to establish thresholds for fire danger days. It might be valuable to discuss potential reasons for this failure and explore whether more sophisticated modeling approaches were considered or could improve predictions.

Data:

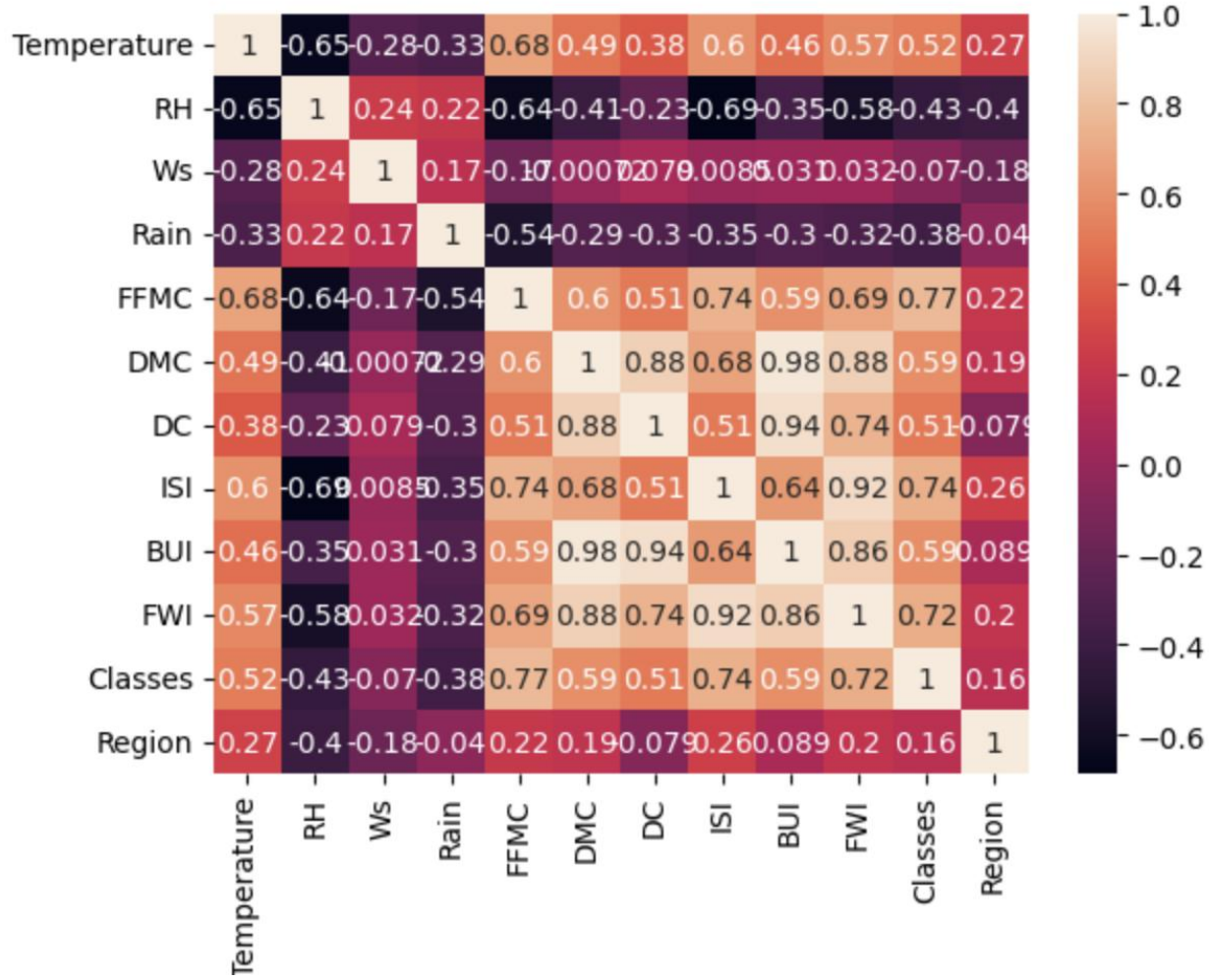
The effectiveness of any model is limited by the quality of the dataset employed for model training. The primary hurdle lies in sourcing an appropriate dataset that encompasses specific features [15]. It has been shown that terrain features, inclinations, gradients, land surface temperatures, humidity levels, wind speed, wind direction, carbon indices, and vegetation indices are effective input features in forest fire prediction [15]. Additionally, data is like fuel to a race car: the larger the data set is, the more representative of the real world it is. And thus, the more accurate it will be.

The Montesano, Portugal dataset selected for this paper was found on Kaggle, a repository of publically available datasets and sourced from *A data mining approach to predict forest fires using meteorological data* [3]**.** Unfortunately, this dataset does not have all of the above features as no such dataset has yet to exist. In the Future Work section there will be a more robust discussion on improving the input data set. This specific dataset, referred to the Portugal Forest Fires dataset, has the following parameters:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "jan" to "dec"
4. day - day of the week: "mon" to "sun"
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

We also did an analysis on all of these features to understand the correlation between them. This is best summarized with this correlation heatmap. As expected, quite a few of the parameters have significant correlation, this could affect the results of the model.



The Algeria dataset includes 244 instances that regroup data of two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. There are 122 instances for each region. The period from June 2012 to

September 2012. The 244 instances have been classified into fire (138 instances) and not fire (106 instances) classes. The dataset includes 11 parameters which are:

1. day', - day of the week: "mon" to "sun"
2. 'month', - 4 months (June, July, August, & September)
3. 'year', - 2012
4. Temperature', ' - temperature noon (temperature max) in Celsius degrees: 22 to 42
5. RH', ' - Relative Humidity in %: 21 to 90
6. Ws', - Wind speed in km/h: 6 to 29
7. 'Rain', ' - total day in mm: 0 to 16.8
8. FFMC', - Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5
9. 'DMC', ' - Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9
10. DC', ' - Drought Code (DC) index from the FWI system: 7 to 220.4
11. ISI', - Initial Spread Index (ISI) index from the FWI system: 0 to 18.5
12. 'BUI', - Buildup Index (BUI) index from the FWI system: 1.1 to 68
13. 'FWI', ' - Fire Weather Index (FWI) Index: 0 to 31.1

The original plan was to combine data set 2 to data set 1 so that there is more training and test data which may improve the overall accuracy. However when combining the two data sets some of the parameters (columns) were removed. Hence the resultant data set was less robust than the original two data sets. Therefore we decided not to pursue the combined data set and ran and trained different models on the two datasets. The Montesano, Portugal Data Set has 12 parameters and 516 instances and the Algeria Data Set has 13 parameters and 243 entries.

The data preprocessing is covered in the section below. Per standard machine learning practice, we split the data 70:30 into training: testing. This allows the model to learn from 70% of the data and evaluate its performance on the remaining 30%. Evaluation metrics like accuracy and precision help assess the model's effectiveness. We are using test accuracy as a metric of our results.

The data within this table set could become less predictive to current climates due to escalating global temperatures. Consequently, the vegetation has become significantly drier,

creating an environment more susceptible to ignition incidents [17]. Climate change has also led to a proliferation of unpredictable weather patterns, such as unexpected summer rain amidst humid conditions, which may differ from the patterns ML models can detect from the two datasets. Data could quickly become outdated due to unpredictable weather patterns. Therefore, collecting near real time data becomes paramount for the accuracy of the prediction.

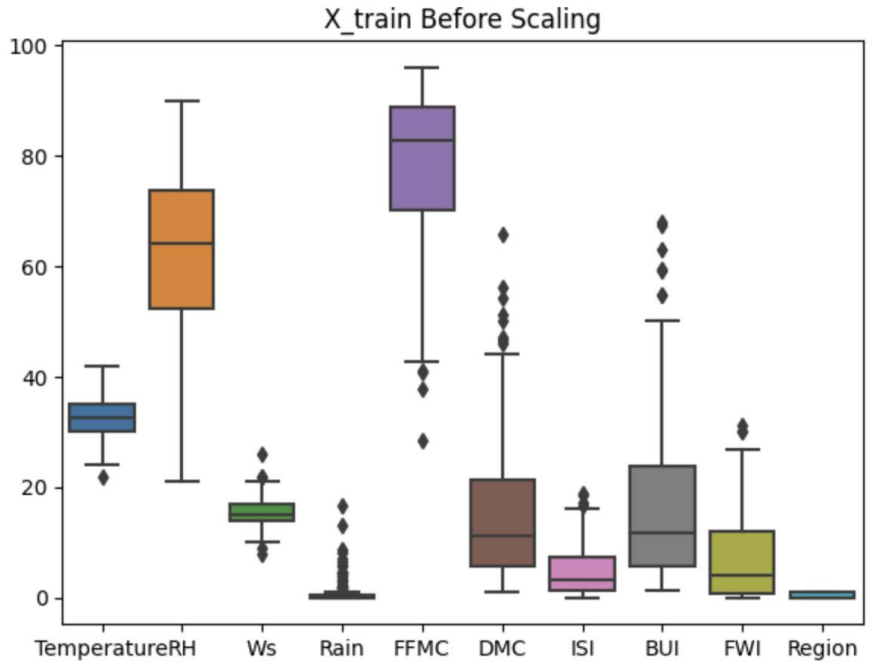
Methodology:

Data processing

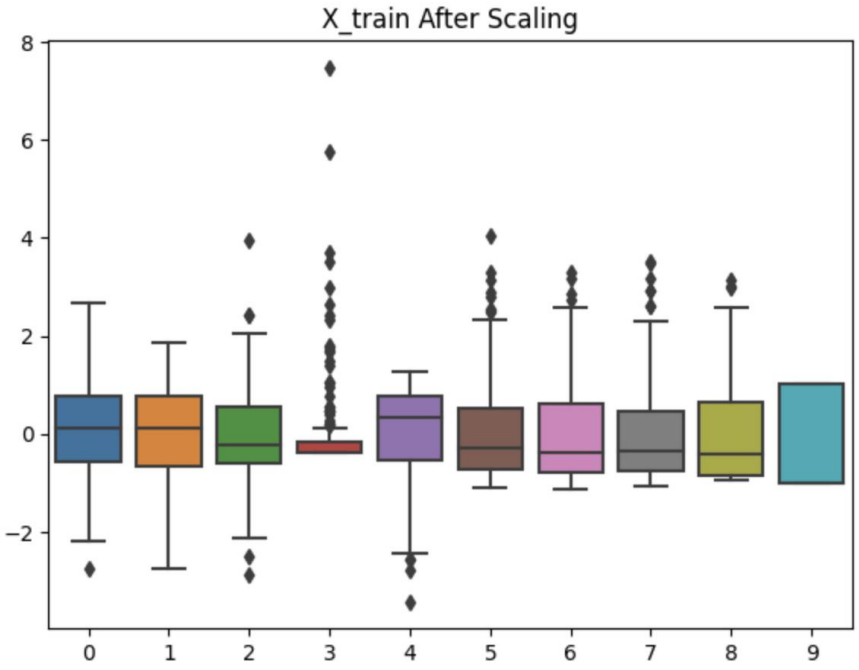
In order to begin training our models, we need to take a few steps to pre-process our data. In the Algerian dataset we begin by removing all null values. Then, in both datasets all of the categorical variables are represented numerically. This includes:

- Representing months and days from strings to points on a unit circle
- Transforming the classification of Fire or No Fire into Binary values

Finally, we normalized all of the data with the scikit-learn standard scaler function. These results are best summarized in the following figures.



Before: Data is unstructured as a result it greatly hinders the accuracy since all the values are skewed left or right



After: Data is processed using Scalar preprocessing function eliminating the majority of the skew making the data easier to read and correspond to other factors on the graph

Models

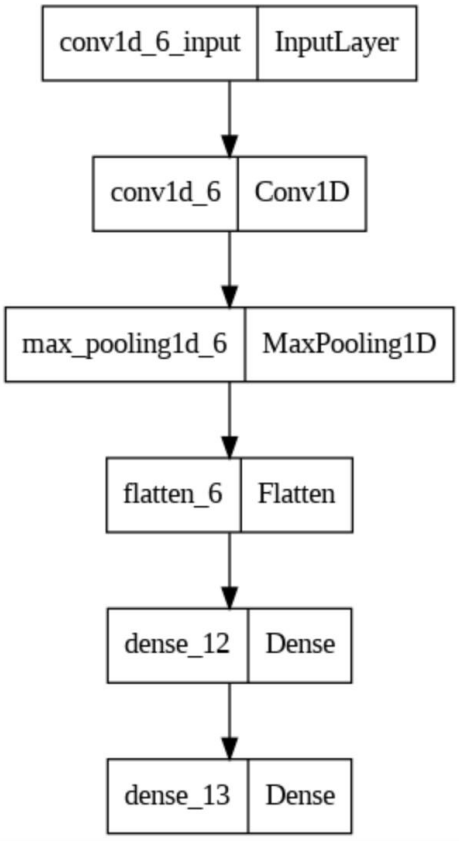
Portugal Dataset

For the Montesano, Portugal data set, we compared the following ML models: Logistic Regression, K-Nearest Neighbors Classifier, Random Forest Classifier, Decision Tree Classifier, and MLP Classifier. We thoroughly compared all the accessible machine learning models to identify the most fitting one that offers superior accuracy.

The utilization of CNN has significantly elevated the landscape of deep learning algorithms, profoundly benefiting various other domains. In order to achieve higher accuracy we used a Convolutional Neural Networks. Specifically we used one convolutional layer in a Cnn

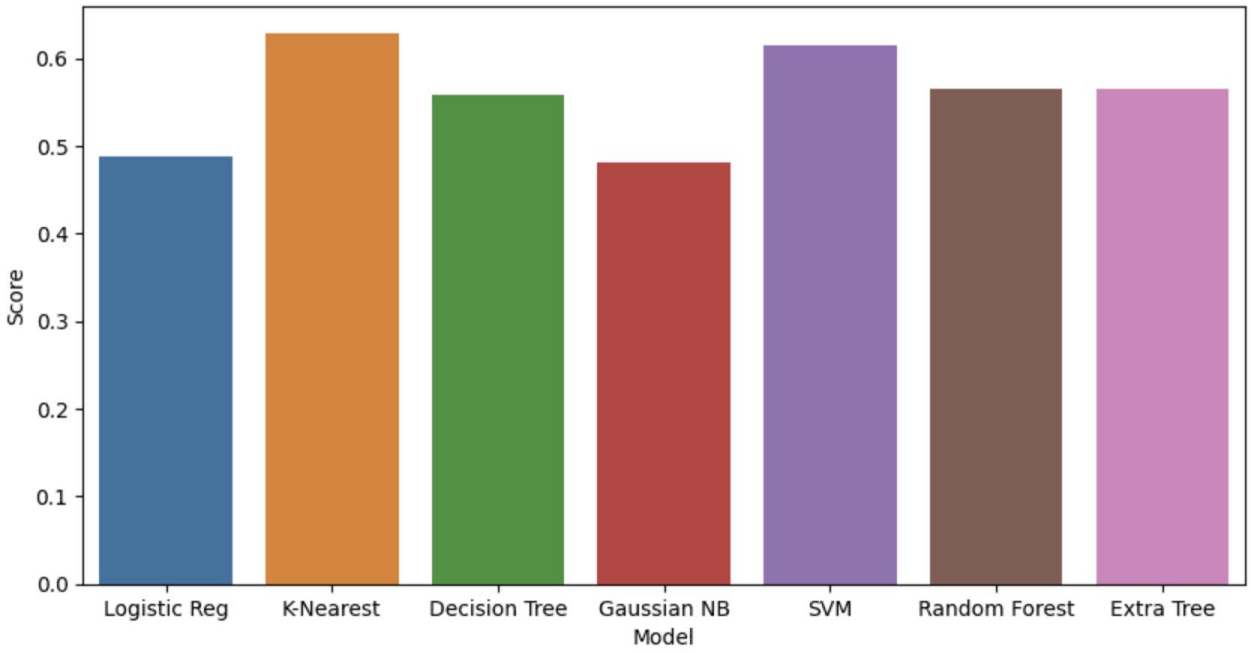
model which analyzes the data on a one dimensional aspect. The following table shows 128 parameters which are manipulated into a trainable data set.

```
[ ]   1 Cnn_model.summary()

    Model: "sequential_6"
    _________________________________________________________________
     Layer (type)                Output Shape              Param #
    =================================================================
     conv1d_6 (Conv1D)           (None, 8, 32)             128

     max_pooling1d_6 (MaxPooling  (None, 4, 32)            0
     1D)

     flatten_6 (Flatten)         (None, 128)               0

     dense_12 (Dense)            (None, 100)               12900

     dense_13 (Dense)            (None, 10)                1010

    =================================================================
    Total params: 14,038
    Trainable params: 14,038
    Non-trainable params: 0
    _________________________________________________________________
```



The flow Chart on the left shows each step in the process taken to compress the Data into a trainable format for the CNN model to use for prediction analysis.

	Type of model	# of parameters	# of epochs	Accuracy
Data Set 1	ML - Logistic Regression	12	N/A	49%
Data Set 1	ML - KNeighborClassifier	12	N/A	63%
Data Set 1	ML - RandomForestClassifer	12	N/A	58%
Data Set 1	ML - DecisionTreeClassifier	12	N/A	53%
Data Set 1	ML - MLPClassifier	12	N/A	51%
Data Set 1	DL - CNN	12	200	68%
Data Set 1	DL RNN	12	200	57%



Montesano, Portugal data set accuracy chart

Out of all the models the best accuracy we obtained is 68% using the Convolutional Neural Networks model. We may be seeing lower accuracy due to a few factors. The majority of fire area (the feature we are trying to predict) are clustered around zero, despite the range being 0.00 to 1090.84 ha. As such, the authors suggest transforming the data with the inverse of ln(x+1) to account for this skew. Additionally, it was originally a regression problem and we changed it to be a classification problem. Thus, we considered a separate dataset that was formed specifically for classification tasks.

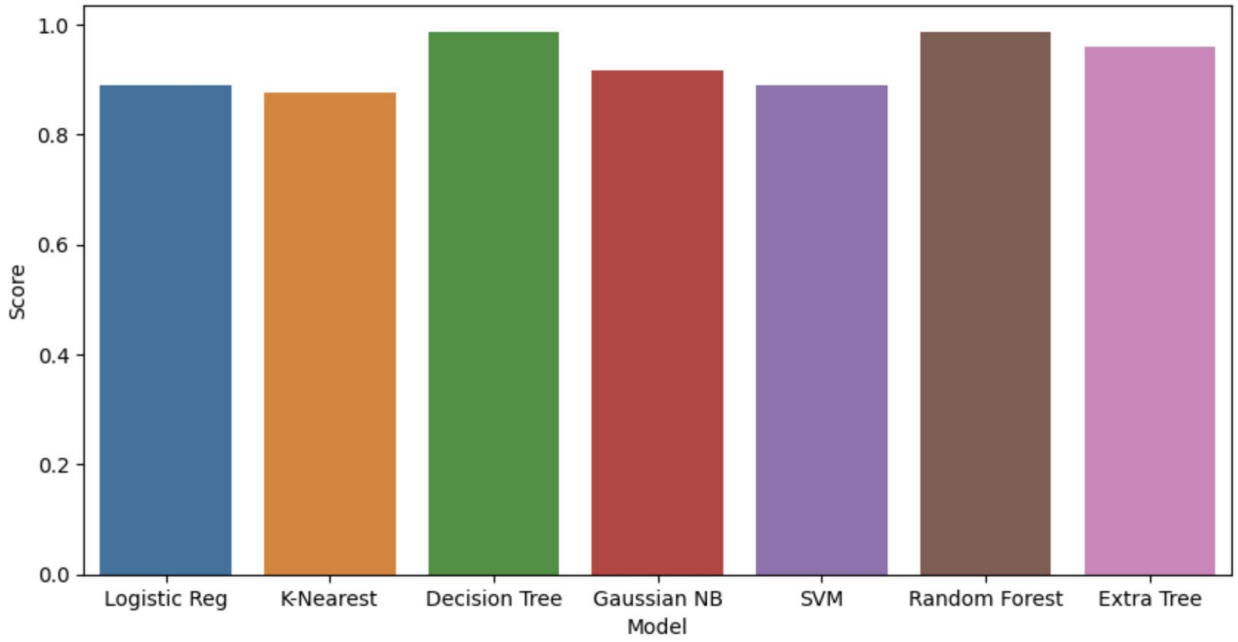
Algeria Dataset

Using the Algeria data set we decided to train a few Machine Learning models, Standard Scalar, Linear Regression, and Logistic Regression, and Deep Learning Models, Convolutional Neural Network and Recurrent Neural Networks.

Algeria **data set accuracy chart**

Data Set 2	ML - LogisticRegression	14	N/A	95%
Data Set 2	ML - K-Nearest Neighbor	14	N/A	94%
Data Set 2	ML - Decision Tree	14	N/A	98%
Data Set 2	ML - Gaussian NB Model	14	N/A	96%
Data Set 2	ML - SVM	14	N/A	92%
Data Set 2	ML - Random Forest	14	N/A	98%
Data Set 2	ML - Extra Tree	14	N/A	95%
Data Set 2	DL - CNN	14	200	96%

I decided to change the types of models by adding SVM and Gaussian NB because I believe that it is more relevant to the data than MLP Classifier. With the Algeria data set we were able to achieve the highest accuracy of 98% with the Machine Learning Models, Decision Tree and Random Forest.



Conclusion :

Analysis

Even though the Algeria data set achieved a higher accuracy, the number of entries in the Algeria data set was less than half of the Montesano, Portugal data. Therefore, we will go with the Montesano, Portugal data set. We believe that the more entries that the model trains on, the more realistic the results are. Therefore we need real time topography sensitive data, including local factors to that particular area to be fed into our algorithm if this were to be implemented. The real time data will be the best fuel for our AI engine.

Future work

In the future we plan to collect dataset from various sources and combine it to a grand dataset which could potentially yield higher accuracy with the standard parameters like weather. Integrating these results with location sensitive data in our model, we expect even higher accuracy predictions. By collecting datasets from recent fires like Santa Cruz CZU lightning complex and Lahaina fire, Maui, Hawaii, we can train our model for more realistic outcomes.

Some of the local specific datasets include :

- Land Surface Temperature(LST)
- Curvature of the Earth
- Low orbit satellite
    - satellite imagery
    - weather patterns & temperature fluctuations
- historical fire data.

The current methods of wildfire detection, although valuable, possess limitations rooted in cost, human labor, and technological constraints. In addition with the rise of Global warming the current datasets start to become less relevant. Embracing AI as a predictive tool offers a promising avenue to revolutionize wildfire detection. In addition through the creation of a network of drones we can accurately collect many more entries and parameters that can accurately describe the topography that can yield more precise predictions of wildfires,

applicability and local factors. By exploiting AI's capacity to decipher intricate relationships within data, we can aspire to mitigate the destructive impact of wildfires, thus safeguarding both human lives and the delicate balance of wilderness ecosystems.

REFERENCES

1. “Data.” *CiteSeerX*, csxstatic.ist.psu.edu/downloads/data. Accessed 9 Sept. 2023.
2. Source: https://archive.ics.uci.edu/ml/datasets/forest+fires
3. P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data.
4. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence,
5. Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December,
6. Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf
7. “Envirovision Home.” *EnviroVision Solutions*, 27 June 2022, evsolutions.biz/.
8. “Ai Wildfire Detection System to Save Lives and Cut Loss: Roboticscats.” *Robotics Cats*, 13 July 2022, roboticscats.com/.
9. GmbH, byteKultur. “20 Years of Automated Early Wildfire Detection.” *Home: IQ FireWatch*, www.iq-firewatch.com/. Accessed 9 Sept. 2023.
10. “Pano Ai.” *AI*, www.pano.ai/. Accessed 9 Sept. 2023.
11. “Earth from Orbit: Tracking Fires from Space.” *National Environmental Satellite, Data, and Information Service*, www.nesdis.noaa.gov/news/earth-orbit-tracking-fires-space. Accessed 9 Sept. 2023.
12. “Silvanet Wildfire Detection: Dryad Networks.” *Dryad*, www.dryad.net/. Accessed 9 Sept. 2023.
13. “Informations et Ressources Scientifiques Sur Le Développement Des Zones Arides et Semi-Arides.” *Analyse Du Bilan Des Incendies de Forêts Dans La Wilaya de Sidi Bel Abbes Durant La Période 2010-2016 - Sécheresse Info*, www.secheresse.info/spip.php?article79279. Accessed 10 Sept. 2023.
14. wfca_teila. “Deadliest Wildfires in U.S. History.” *WFCA*, 23 Aug. 2023, wfca.com/articles/deadliest-wildfires-in-us-history/.
15. Zhang, Guoli, et al. “Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China - International Journal of Disaster Risk Science.”

SpringerLink, Beijing Normal University Press, 19 Sept. 2019, link.springer.com/article/10.1007/s13753-019-00233-1.

16. “Alertcalifornia.” *ALERTCalifornia*, 5 Sept. 2023, alertcalifornia.org/.
17. “3 Reasons Wildfires Are Getting More Dangerous-and 3 Ways to Make Things Better.” *The Wilderness Society*, www.wilderness.org/articles/blog/3-reasons-wildfires-are-getting-more-dangerous-and-3-ways-make-things-better?gad=1&gclid=Cj0KCQjwx5qoBhDyARIsAPbMagDVWwLvT-7zd-eaqrBrivBoRqTYqixpUWpmazVP9lam0sWya-Bl3asaAvVcEALw_wcB. Accessed 17 Sept. 2023.

Acknowledgement :

- Mentor and Advisor: Juyon Lee , B.Sc Computer Science , Stanford University ; M.Sc Computer Science Oxford University
- Consulting Advisors : Kurtis Nelson at Earth Resources Observation and Science (EROS) Center