C951 INTRO TO AI – TASK 3

Josh Gaweda

TABLE OF CONTENTS

A. CREATE A PROPOSAL FOR A MACHINE LEARNING PROJECT BY DOING THE FOLLOWING:

1. DESCRIBE AN ORGANIZATIONAL NEED THAT YOUR PROJECT PROPOSES TO SOLVE.

The intent of this project is to bolster firefighting efforts in the state of California by better understanding how flames start and spread. This solution aims to save the State of California billions and maybe even trillions of dollars in disaster recovery costs due to an effective machine learning fire prediction model.

2. DESCRIBE THE CONTEXT AND BACKGROUND FOR YOUR PROJECT.

Over the past several decades, climate change and deforestation have led to a growing number of forest fires every year. This past year wildfires reached historic numbers in regard to the number of fires and how large they became throughout California and much of the West Coast. In order to best fight these fires we need to understand how they spread and what we can do to best mitigate their damage. The primary aim is to assess the feasibility of utilizing an open source Weather Research and Forecasting(WRF) simulator along with Physics Informed Neural Networks(PINN) to predict fire spread.

3. REVIEW **THREE** OUTSIDE WORKS THAT EXPLORE MACHINE LEARNING SOLUTIONS THAT APPLY TO THE NEED DESCRIBED IN PART A1.

1. Fire Safety Journal: Volume 104**[1]**- This volume of the *Fire Safety Journal* highlights how forest fires are an increasing problem and one of the more devastating disasters all over the world. This edition takes an in depth look at different fire prediction models, data simulation techniques, and different methods for forecasting the occurrence of forest fires.

2. *International Journal of Recent Technology and Engineering***[2] -** This journal details how quickly fires spread and discusses how machine learning and data mining techniques can help to predict the spread of fires. This journal also goes deeper into the examination of high fire risk areas before the fire begins

3. A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems**[3]-** This paper presents a thorough examination of machine learning algorithms for fire detection and prevention and how the threat can be mitigated with proper planning. This paper also discusses the limitations of such models and what we can do to improve the model deficiencies.

3A. DESCRIBE HOW *EACH* REVIEWED WORK FROM PART A3 RELATES TO THE DEVELOPMENT OF YOUR PROJECT.

1. Fire Safety Journal:Volume 104 is the first piece of literature I found on this the topic of machine learning for predicting forest fires. I found this while researching robot disaster recovery in Task 2 and I guess you could say this was my primary inspiration for choosing wildfires prediction as my project proposal. This journal goes over some surface level information on 12 different fire detection algorithms that can be used for predicting wildfires. These algorithms are more geared towards predicting flame intensity and not so much predicting start or spread. I did not end up using any of the algorithms mentioned in this journal, but I did learn quite a bit about the pattern-like nature of wildfires and had my interest piqued on the topic. **[1]**

2. The International Journal of Recent Technology and Engineering mentions the WRF simulator and how other teams are trying to use it throughout the country. This led to the WRF simulator being selected as the platform of choice for our machine learning model. Using an open source platform gives us a lot of flexibility with post deployment support and makes the initial implementation a little bit easier as well. **[2]**

3. A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems talks about several machine learning algorithms including their use of PINNs in their final prediction model. After doing some research I found that these physics inspired neural nets are very good at modeling physical laws, including a flame. Since I had already chosen WRF as my platform, I reviewed the documentation and found that the equation solving engine was easily swappable with my own open source option (NeuralPDE). **[3]**

4. SUMMARIZE THE MACHINE LEARNING SOLUTION YOU PLAN TO USE TO ADDRESS THE ORGANIZATIONAL NEED DESCRIBED IN PART A1.

We would use PINNs to train our fire prediction models inside the WRF simulator. PINNs are a type of partial differential equation solver that normalize the physical information in an equation to improve performance over more conventional differential equation solvers. More specifically will be using an open source PINN called NeuralPDE.jl**[8]**. The WRF simulator is a state-of-the-art open source weather modeling system. The edition of WRF we are working with is called WRF-SFIRE**[7]**.

5. DESCRIBE THE BENEFITS OF YOUR PROPOSED MACHINE LEARNING SOLUTION

Differential Equations are a critical component to understanding the world around us. They are often used to explain well known physical laws in science and engineering, and are ubiquitous in the study of fluid dynamics and heat dispersion. PINNs are built with these types of physical laws in mind so naturally they would be a top pick for fire prediction. PINNs are well documented and so is the WRF simulator we would use to implement the neural network. Implementing WRF with PINNs could help us with early warning systems for fires, identify areas with high fire risk, and help responders with fire control once the flames have already been ignited.

B. DESCRIBE YOUR PROPOSED MACHINE LEARNING PROJECT PLAN BY DOING THE FOLLOWING:

1. DEFINE THE SCOPE OF THE PROPOSED MACHINE LEARNING PROJECT.

In Scope:

- Create a fire forecast system to help with California wildfires.
- Setup of fire forecast system for use throughout California.
- Training for Forest Rangers and California fire prevention professionals on how to use the system.

Not it in scope:

- The system will only attempt to forecast *forest* fires and will not attempt to predict structural fires.
- The system will not be optimized for areas outside of California

2. EXPLAIN THE GOALS, OBJECTIVES, AND DELIVERABLES FOR THE PROPOSED PROJECT.

Goals:

- To lower loss of human life and destruction from forest fires as much as possible.
- Reduce response time to forest fires.
- Assist first responders in slowing fire spread and determining safest escape routes.

Objectives:

- Implement the prediction model by July 2022(before peak wildfire season in CA).
- Provide responders with a measured reduction in fire response time by the end of September (conclusion of primary wildfire season) 2022.

Deliverables:

- Create a fire warning and prediction system that can be used with minimal training by first responders and other fire prevention professionals.
- Provide municipal level support for first year after distribution. After 1 year development and support will be left up to current State level officials in charge of fire prevention.

3. EXPLAIN HOW YOU WILL APPLY A STANDARD METHODOLOGY (E.G., CRISP-DM, SEMMA) TO THE IMPLEMENTATION OF YOUR PROPOSED PROJECT.

The CRISP-DM methodology has 6 steps**[9]** outlined below that that will lead to a successful implementation of our solution:

Business understanding

The state of California needs a reliable wildfire prediction model to save lives and prevent damages. The benefit of implementing this solution far outweighs the minimal cost to taxpayers(outlined in the cost analysis chart located in B5). As noted in the cost-analysis, the primary cost will be paying the employees as the data is provided free from NASA.

By using NASA's data and feeding it to our PINN we can create a working prediction model for fire prevention professionals in California. Once the model is in place it will be ready for distribution and training through the state of California. Further support and development will be provided for the first year. At the conclusion of the first year of support, the project will be handed over to the new state level officials in charge.

Data understanding

The data being used is in HDF-EOS5 format. This format is the recommended standard for use in Earth Science Data Systems and approved by NASA since 2007**[9]**. This data can be easily collected on a daily basis from NASA's data repository located at the link in section D1. HDF-EOS files are multi-object file formats so there are many ways to work with them. For each object inside these files there are predefined labels for type, amount, and data dimensions that describe what is inside the object. **[9]** Our simulation software can automatically interpret and identify these labels to produce multi-dimensional images that are also geolocated. Typically, these two data types would need to be loaded separately and then merged, however the HDF-EOS format already has the two combined. With this data being sourced from NASA there are no quality concerns. Additional information on this file format can be found at the following link: https://hdfeos.org/

Data preparation

The data will not require much preparation as the weather simulator can natively open the HDF-EOS5 file format.

Modeling

We will use an open-source Physics Informed Neural Net(PINN) located on github inside the NeuralPDE repository. Figure 1 is an example of how our prediction model might be derived using a PINN. **[8]**

```
using NeuralPDE, Flux, ModelingToolkit, GalacticOptim, Optim, DiffEqFlux

@parameters x y
@variables u(..)
Dxx = Differential(x)^2
Dyy = Differential(y)^2

# 2D PDE
eq  = Dxx(u(x,y)) + Dyy(u(x,y)) ~ -sin(pi*x)*sin(pi*y)

# Boundary conditions
bcs = [u(0,y) ~ 0.f0, u(1,y) ~ -sin(pi*1)*sin(pi*y),
       u(x,0) ~ 0.f0, u(x,1) ~ -sin(pi*x)*sin(pi*1)]
# Space and time domains
domains = [x ∈ IntervalDomain(0.0,1.0),
           y ∈ IntervalDomain(0.0,1.0)]
# Discretization
dx = 0.1

# Neural network
dim = 2 # number of dimensions
chain = FastChain(FastDense(dim,16,Flux.σ),FastDense(16,16,Flux.σ),FastDense(16,1))

discretization = PhysicsInformedNN(chain, GridTraining(dx))

pde_system = PDESystem(eq,bcs,domains,[x,y],[u])
prob = discretize(pde_system,discretization)

cb = function (p,l)
    println("Current loss is: $l")
    return false
end

res = GalacticOptim.solve(prob, Optim.BFGS(); cb = cb, maxiters=1000)
phi = discretization.phi
```

Figure 1

Analysis of the resulting data would look something like Figure 2 below**.[8]**

```
xs,ys = [domain.domain.lower:dx/10:domain.domain.upper for domain in domains]
analytic_sol_func(x,y) = (sin(pi*x)*sin(pi*y))/(2pi^2)

u_predict = reshape([first(phi([x,y],res.minimizer)) for x in xs for y in ys],(length(xs),length(ys)))
u_real = reshape([analytic_sol_func(x,y) for x in xs for y in ys], (length(xs),length(ys)))
diff_u = abs.(u_predict .- u_real)

using Plots
p1 = plot(xs, ys, u_real, linetype=:contourf,title = "analytic");
p2 = plot(xs, ys, u_predict, linetype=:contourf,title = "predict");
p3 = plot(xs, ys, diff_u,linetype=:contourf,title = "error");
plot(p1,p2,p3)
```

Figure 2a

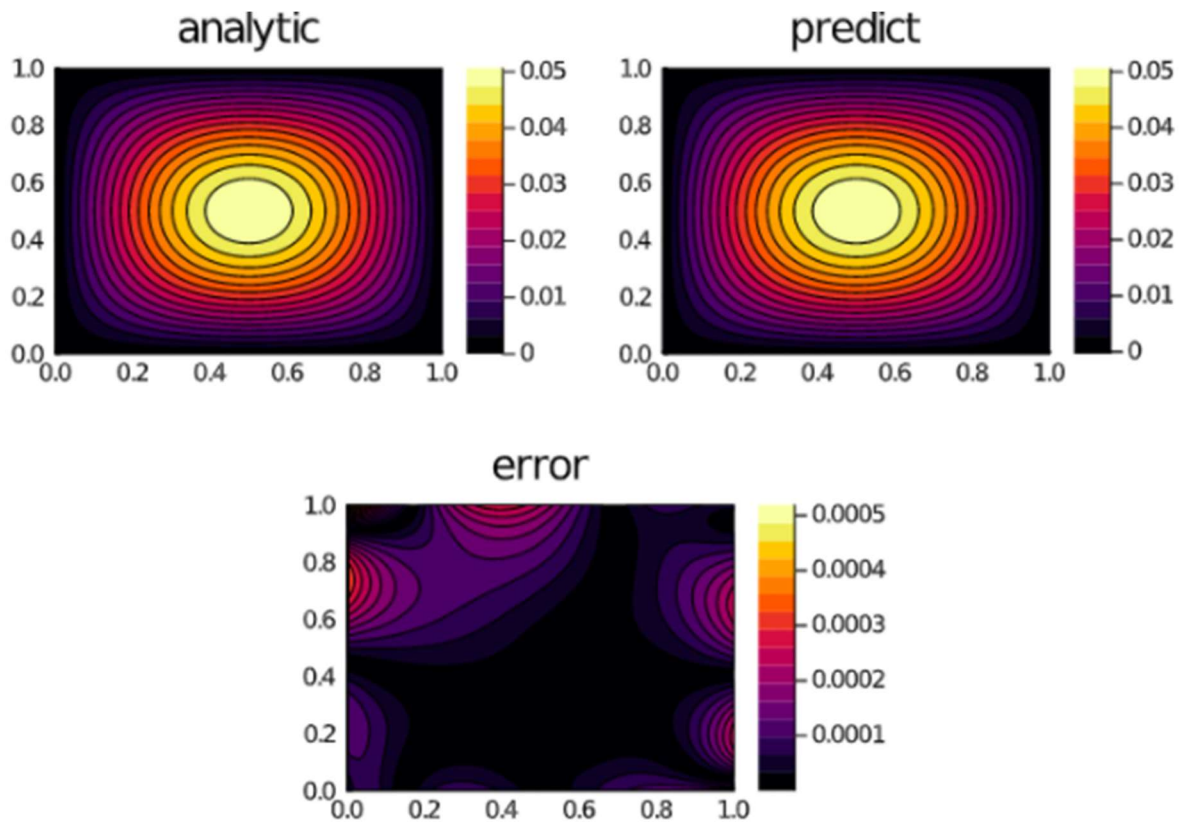

Figure 2b

Evaluation

This machine learning model seems to fit our need predicting wildfires in California. This solution is cost effective and we believe it can be effective in the field. If unseen issues arise, we can pivot to another machine learning solution inside the NeuralPDE package, but that is not likely.

Deployment

Standard training sessions will be provided via Zoom in order to maximize the number of professionals we reach at one time and reduce training costs. Our deployment will consist of working closely with individual municipalities throughout California, that are deemed high risk. This rollout will take place beginning in February 2022 and last for 4 months. After this initial rollout period, our Analysts and Engineers will offer support for 1 year as needed.

4. PROVIDE A PROJECTED TIMELINE FOR THE PROPOSED PROJECT, INCLUDING THE START AND END DATES FOR *EACH* TASK.

Task	Timeline
Being fully informed on the most common causes of fires and how to model fires in a simulation environment.	Project Begin 2021 – Project End 2022 (Est. 1 Year)

Estimated Project Start Date	March 2021
Implement machine learning model with WRF simulator.	March 2021 – May 2021
Collecting, interpreting, preparing, and sharing our collected data with other fire prevention professionals	May 2021 – September 2021
Distributing and training users on the fire forecasting solution in preparation for fire season 2022.	Feb 2022 – May 2022
Provide support for 1 year after training/distribution.	Feb 2022 – May 2023

5. LIST RESOURCES (E.G., HARDWARE, SOFTWARE, WORK HOURS, THIRD-PARTY SERVICES) AND *ALL* ASSOCIATED COSTS NEEDED TO IMPLEMENT THE PROPOSED SOLUTION.

Resource	Quantity	Est. Cost/Year	Years Needed	Total
NeuralPDE.jl	Open-Source	0	2	0
WRF-SFIRE	Open-Source	0	2	0
Data Collectors	4	40,000	1	160,000
Data Analysts	4	75,000	1.5	150,000
Data Scientist/Engineer	5	100,000	2	100,000
Workstation	13	4,000	1	52,000
Data Server	1	10,000	1	10,000
		229,000		647,000

NeuralPDE.jl is available here**[7]:** https://github.com/SciML/NeuralPDE.jl

WRF-SFIRE is available here**[6]:** https://github.com/openwfm/WRF-SFIRE

6. DESCRIBE THE CRITERIA THAT YOU WILL USE TO EVALUATE THE SUCCESS OF THE PROJECT ONCE IT IS COMPLETED.

- Fully implement the prediction model by July 2022
- Provide responders with a measured reduction in fire response time by the end of September 2022.
- Complete 1 year of supportive training for all California fire prevention professionals.

C. DESCRIBE THE PROPOSED MACHINE LEARNING SOLUTION YOU WILL USE TO ADDRESS THE ORGANIZATIONAL NEED IDENTIFIED IN PART A1 BY DOING THE FOLLOWING:

1. IDENTIFY THE HYPOTHESIS OF THE PROPOSED PROJECT.

The hypothesis is that our machine learning model can effectively reduce death and destruction due to wildfires by predicting when and where they will most likely start and spread.

2. IDENTIFY THE MACHINE LEARNING ALGORITHM(S) (I.E., SUPERVISED, UNSUPERVISED, OR REINFORCEMENT LEARNING) YOU WILL IMPLEMENT IN YOUR PROPOSED SOLUTION.

The NeuralPDE PINN algorithm is Supervised.

A. JUSTIFY THE SELECTION OF THE ALGORITHM IN PART C2. INCLUDE **ONE** ADVANTAGE AND **ONE** LIMITATION OF THE SELECTED MACHINE LEARNING METHOD.

The NeuralPDE was a natural advantage for the purpose of fire prediction as PINNs are designed to model physical laws.**[3]** Anything from fire, to swarm behavior, to virus infection, can be modeled with PINNs. The primary limitation is the amount of data that will need to be stored and since we have purchased a large data server, we should not run into our upper data limit during this time frame.

3. DESCRIBE THE TOOLS AND ENVIRONMENTS THAT WILL BE USED TO DEVELOP THE PROPOSED MACHINE LEARNING SOLUTION, INCLUDING ANY THIRD-PARTY CODE.

- Provided Workstations with data server access.
  - 64GB RAM
  - i9 Processor
  - Data server houses up to 5 PB of data
- Python and Julia
- Juno IDE
- NeuralPDE.jl for Physics Informed Neural Network(PINN) equation solving. Open source, available on github.
- WRF-SFIRE for fire spread simulation. Opensource meteorological simulation software. Also available on github and the WRF website.

4. EXPLAIN THE PROCESS YOU WILL USE TO MEASURE THE PERFORMANCE OF YOUR PROPOSED MACHINE LEARNING SOLUTION.

In neural networks there are rules for measuring the correct classification of something. **[4]**
If we let $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$ be the vector consisting of the predicted class labels with $\hat{y}_1 \in 1, \ldots, R$, the correct classification rate(CCR) can then be expressed as:

$$CCR = 1/n \sum_{i=1}^{n} \delta(y_i, \hat{y}_i)$$

In this formula , δ, is an indicator variable such that $\delta(y_i, \hat{y}_i) = 1 \text{if } y_i = \hat{y}_i$ and otherwise zero. This formula is based on training data, so it will often give an exaggerated view of the outcome. Because of this, it is often a good idea to use the formula noted below as the secondary classification model:

$$\mathrm{CCR}_{\mathrm{test}} = 1/\mathrm{n}' \sum_{\mathrm{i}=1}^{\mathrm{n}'} \delta\left(\mathrm{y}'_{\mathrm{i}}, \widehat{\mathrm{y}}'_{\mathrm{i}}\right)$$

[5]

D. DESCRIBE THE DATA FOR YOUR PROPOSED PROJECT BY DOING THE FOLLOWING:

1. IDENTIFY THE SOURCE(S) OF THE DATA FOR YOUR PROPOSED PROJECT.

Our data source will primarily be NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) satellite armed with specialty meteorological sensors. This data source is freely available to the public for uses such as this.

Daily NASA data can be found at the following link**[5]:**
https://lpdaac.usgs.gov/product_search/?sort=title&view=list&temporal_resolution=%3C+Daily

2. DESCRIBE THE DATA COLLECTION METHOD.

For our prediction model, we will use data from MODIS Active Fire system. This data describes features on land that can be used to predict fire risk. There are various data sets included with MODIS, including satellite imagery which we can use for our detection models as well.

A. DISCUSS **ONE** ADVANTAGE AND **ONE** LIMITATION OF THE DATA COLLECTION METHOD DESCRIBED IN PART D2.

The main advantage of the data is that it is trusted by most national organizations associated with wildlife and meteorological causes. We can trust the data we are getting is from the most trusted source available and is getting looked at by the best minds inside our national borders. The main limitation I would say is that there is only one data source. With anything, it's nice to have multiple sources for your data to smooth out any bias that might be present. Fortunately, I think we are safe with NASA data as it is the most trusted source for data in the HDF-EOS5 format.

3. EXPLAIN HOW YOU WILL PREPARE YOUR DATA FOR USE BY THE MACHINE LEARNING ALGORITHM(S) FROM PART C2 FOR YOUR PROPOSED PROJECT, INCLUDING DATA SET FORMATTING, MISSING DATA, OUTLIERS, DIRTY DATA, OR MITIGATION OF OTHER DATA ANOMALIES.

Fortunately, several entities already provide open access to their own data viewing format for MODIS that we can use for free. The United State Geological Survey(USGS) provides their own interface that we can use for prediction here: https://lpdaac.usgs.gov/tools/usgs-earthexplorer/

To ensure that we have the most complete picture as possible, in regard to outliers and missing data points, we will cross-reference multiple USGS interfaces to smooth out any noise.

4. DESCRIBE BEHAVIORS THAT SHOULD BE EXERCISED WHEN WORKING WITH AND COMMUNICATING ABOUT SENSITIVE DATA IN YOUR PROJECT.

All data will be open source and would be provided to any third party that asked. Data integrity will be more important than anyone gaining unauthorized access.

E. ACKNOWLEDGE SOURCES, USING IN-TEXT CITATIONS AND REFERENCES, FOR CONTENT THAT IS QUOTED, PARAPHRASED, OR SUMMARIZED.

1. Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal, 104*, 130-146. doi:10.1016/j.firesaf.2019.01.006

2. Predicting forest fires using supervised and Ensemble machine learning algorithms. (2019). *International Journal of Recent Technology and Engineering, 8*(2), 3697-3705. doi:10.35940/ijrte.b2878.078219

3. Abid, F. A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems. *Fire Technol* **57,** 559–590 (2021). https://doi.org/10.1007/s10694-020-01056-z

4. Correct classification. (n.d.). Retrieved March 01, 2021, from https://www.sciencedirect.com/topics/computer-science/correct-classification

5. NASA DATA. (n.d.). Retrieved March 02, 2021, from https://lpdaac.usgs.gov/product_search/?sort=title&view=list&temporal_resolution=%3C%2BDaily

6. Openwfm. (n.d.). Openwfm/wrf-sfire. Retrieved March 02, 2021, from https://github.com/openwfm/WRF-SFIRE

7. SciML. (n.d.). SciML/NeuralPDE.jl. Retrieved March 02, 2021, from https://github.com/SciML/NeuralPDE.jl

8. Crisp-DM. (2021, February 27). Retrieved March 02, 2021, from https://www.datascience-pm.com/crisp-dm-2/

9. The HDF-EOS Tools and Information Center. (n.d.). EOS tools and Information Center. Retrieved March 02, 2021, from https://hdfeos.org/