Alexander He
hea2@rpi.edu
Class of 2022

Remember that it is imperative that you submit all required assignments, as noted as participation in the away semester course. Failure to do so will result in a U for the away semester requirement. A "U" will impede on your expected graduation.

Due: No Later than May 10th (NO EXCEPTIONS)

ILE-Research: Construct a 2-3 page paper that summarizes the following:

- The scope of your research experience—supervising professor (include full name and contact info at the end of paper), topic and degree of participation on the project, length of assignment.
- Discuss the impact on your professional & educational goals.
- Discuss any relevant employable skills derived from your experience that can be potentially listed on a resume.
- Provide ample details about the highs and lows of semester away experience.

- The scope of your research experience—supervising professor (include full name and contact info at the end of paper), topic and degree of participation on the project, length of assignment.

My research experience was an individual (non-group) project with the end goal of writing a paper and submit it to a conference. My advising professor is Prof. Thilanka Munasinghe. His role has been to hold weekly meetings with me to discuss my progress and give me guidelines. The planned start and end dates of my project are February 8 through May 14. However, since I am submitting a paper to the IEEE DSAA'2021 conference for the Research & Applications Track, I would consider the submission deadline, May 23, as the end of my project.

My original proposal was to research how wildfires and other fire emissions impact air quality, healthcare costs, and the economy. I intended to use spatiotemporal datasets for carbon emissions from fires, particulate matter, climate variables/indices, population, death counts of respiratory diseases, and associated healthcare costs. The final deliverable included in the proposal was to write and submit a research paper to a conference.

After extensive reading of related publications, creating data visualizations, and performing exploratory data analysis (EDA), I eventually decided to narrow the focus to make my project a regression modeling study with death counts as the output variable. I was particularly inspired by the paper "Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables." The authors used random forest from SciKit Learn to classify cholera outbreaks, with monthly lag values of climate variables to represent their lagged effects on cholera incidence. I took a similar approach using random forest, except I performed regression rather than classification, and incorporated fire emissions and biosphere fluxes from Global Fire Emissions Database (GFED), particulate matter from Atmospheric Composition Analysis Group, and climate variables/indices from NOAA Monthly U.S. Climate Divisional Database (nClimDiv).

As a result of my change of plans and the lower accessibility of healthcare data, I could not find a dataset for healthcare costs and analyze the impact on healthcare costs and the economy. Fortunately, I was able to use all other datasets in my original proposal.

- Discuss the impact on your professional & educational goals.

I am pursuing a career related to Data Science and Data Analytics. I am also applying for graduate school programs in Data Science and Computer Science. This project will undoubtedly add to my graduate school applications. If a conference accepts my paper, it would add to my applications even more.

- Discuss any relevant employable skills derived from your experience that can be potentially listed on a resume.

This project would be the first time I applied skills and knowledge I gained from advanced computer science courses at RPI, especially *Data Science* and *Machine Learning from Data*, to a project outside of courses. This project is also the first time I used an open-source package (SciKit Learn) for machine learning. I also became familiar with processing gridded spatiotemporal data using shapefiles and open-source packages such as SciPy, geopandas, and rasterio. Besides listing the use of these libraries, I may add something along the lines of "writing a research paper" to my skills as well.

- Provide ample details about the highs and lows of semester away experience.

Due to the gridded nature of the spatiotemporal datasets, I wrote code to aggregate the data by county based on shapefiles of US county boundaries. I would consider this the most time-consuming part of my project, especially since the data sources provide their data in different formats, so I needed to handle each dataset differently. However, it was a great application of the knowledge I gained from the *Introduction to Algorithms* course for writing efficient algorithms. I also gained experience with the rasterio and geopandas packages, which assist in working with spatiotemporal data.

Since this was my first project involving data science outside of my courses, I was inexperienced with data science/machine learning applications. I had to spend a lot of time reading publications and wrangling with data to come up with my proposal, and also refine my project on the fly. The most significant modification I made to my project was deciding to narrow the focus to a regression modeling study using random forest, as mentioned before. This modification happened around April 10, which is quite late into my project, so the past month has been quite intense. I performed cross-validated feature elimination and hyperparameter testing to optimize the R^2 value of the model based on the available datasets. The features in consideration are composed of all datasets included in the study, with 1-2 month lagged values for GFED datasets, PM2.5, and nClimDiv datasets.

After doing most of the feature elimination, I performed Spearman's Rank Correlation with all datasets and the appropriate lagged values. I should have done this before feature elimination, as it would have narrowed down features to eliminate and sped up the process overall. I now understand how useful it is to look for correlations between features as part of EDA before model building occurs.

This project is also the first time I am writing a paper to submit to a conference. Besides reading publications applicable to this project, I have substantial experience with reading/studying research publications before this project. Thus, drafting my paper was not too daunting. However, my study takes a novel exploratory approach, which calls for extensively discussing my study's limitations and potential relationships with other publications (in the Discussion section). Thus, I needed to spend additional time researching publications for this purpose.

Advising professor:
Prof. Thilanka Munasinghe
munast@rpi.edu