My Journey Through the Harvard Forest Data Archive for the 2021 Annual Harvard Forest Ecology Symposium

Iulia Iordanescu, Acton Boxborough Regional HS, 24iordanescui@abschools.org

How did I end up communicating with researchers associated with Harvard Forest?

I found out about the Harvard Forest (HF) Ecology Symposium from Mr. Dempsey's note on ABHS Research Club. From there I did some research on the symposium web page, and I discovered that Harvard Forest (a 4000 acre laboratory & classroom research entity established by Harvard University) has a very rare feature: they provide open access to a large dataset that can be discovered by searching in multiple ways (general search, by investigator, keyword, year, taxon, ID number, Research Topic and many others). Because access to data is open, I could briefly skim and find a few datasets that looked promising – they had an interesting and relevant topic, and data looked relatively simple to work with (tabular data with tens of columns). The next step was to contact HF staff (some of them are academic researchers, although I did not know this at that time) and introduce myself to them. I mentioned my HS and the fact that this project was suggested by "Mr. Dempsey who is my science club teacher". The response was very positive, and they encouraged me to go ahead. They also encouraged me to contact the data authors and ask for usage permission. I followed their advice and contacted the "owners" for 3 datasets, but only 2 responded, so I ended up with these two datasets:

- HF120: Tree Seed Dispersal in Hemlock Removal Experiment at Harvard Forest 2005, Dr. Aaron Ellison

- HF213: Forest Inventory for Tree Demography and Carbohydrate Reserves at Harvard Forest 2009-2011, Dr. Michael Dietze

My goal was to improve my knowledge of Python and Machine learning by working on real data and HF Symposium and web-site provided the perfect opportunity for this. Both Dr. Ellison and especially Dr. Dietze were very welcoming, and they helped me to understand the questions they were trying to answer with the data they collected. For example, Dr. Dietze said he is not merely interested in predicting tree mortality using tree diameter measurements, he also wants to understand whether the mortality dependency is linear on j shaped. This was a very helpful guidance point, and it further pushed me to extend the analysis and understand which features were important for mortality prediction.

I presented my work as a poster with the title “Tree Mortality prediction and Tree Seed Dispersal modelling using Machine Learning“ on March 16, the first day of Harvard Forest Ecology Symposium 2021.

Tools I used

I used machine learning Python tools and tutorials that are available free on the internet.

Data science:

- **scikit-learn: is an excellent library for Machine Learning in Python**

- **https://towardsdatascience.com: is a very good learning tool, with many free detailed tutorials for classification and regression (and also for feature importance analysis).**

- **https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn**

Compute environment:

- **https://www.docker.com/101-tutorial: my father is a cloud ML engineer and scientist and he helped me here**

- **https://www.dataquest.io/blog/jupyter-notebook-tutorial/: a powerful tool for interactively developing and presenting data science projects.**

What I gained from the experience

Probably the most important experience I gained is the live “interview” I had during my presentation, when real academic researchers (university professors and students) came by my poster and asked me questions. Although it was my work, and I knew what I did, the experience was intimidating and the discussions were exhausting because they knew the science part really well, and most of their questions were about aspects I never thought about before. While for some questions I did not have a clear answer because of my lack of knowledge, others were actually very helpful because by explaining my answer I clarified my understanding and thus I felt more advanced just because I could reorganize my thoughts and I could see the data and its processing from new angles.

Other valuable lessons I learned:

- **Deadlines are extremely difficult when you have to do work you have never done before. Things that I imagined would take 30 minutes ended up taking a whole day.**

- **Data science is very difficult and requires advanced knowledge of math. I was able to finish my analysis, but I will need to learn a lot more before I could think of doing a real academic research project.**

- **My Python knowledge improved fast when I had to process real data. Programming in Python is difficult, but doing data analysis using a good and popular library like scikit-learn is achievable because for many problems there are already published solutions available on community driven web sites like https://stackoverflow.com/ .**