1. Dataset:

I'm inspired by the paper *Tackling Climate Change with Machine Learning* [1], and I'm especially interested in automating afforestation. I want to work on a project related to afforestation, so I choose Forest Cover Type Dataset [2]. This dataset can be used to predict the forest cover type given the surroundings. This technique can be used to determine what species to plant in a particular area to increase the climate impact of afforestation.

2. Methodology :

a) Data Preprocessing: There are 581,012 instances in total and 12 measures of attributes. I plan to follow the classification in Data Set Description [3], first 11,340 records used for training data subset, next 3,780 records used for validation data subset, and last 565,892 records used for testing data subset. For 10 quantitative features, since their ranges differ greatly and most of their distributions are not normal distribution [4], I plan to normalize the data using min-max normalization. The 2 qualitative features have already been encoded as 0 or 1 for "absence" and "presence." I plan to add 2 new features, Distance_To_Hydrology (calculated from Horizontal_Distance_ To_Hydrology and Vertical_Distance_To_Hydrology) and Average_Hillshade (calculated from Hillshade_9am, Hillshade_Noon, and Hillshade_3pm.) It might also be useful to consider the correlations between wilderness type and other features and add more new features with respect to them.

b) Machine Learning Model: This project is to predict the tree cover type based on the surrounding environment. It is a classification problem, so KNN, support vector machine or neural network might be useful models for this project. KNN is the most intuitive algorithm for me since it is natural to predict a tree cover type based on the surrounding tree types. However, KNN is probably too slow for this dataset. After we learn more about classification methods, I will decide which algorithm to use. I also need to plot the data to see whether they are linearly separable.

c) Evaluation Metric: Confusion matrix is useful to this project since there is an imbalance in the dataset [5]. For example, the number of records of Spruce-Fir is 211,840, while the number of records of Cottonwood/Willow is only 2,747. AUC can also be used as an evaluation metric.

3. Application:

I only have a very brief idea about my application and might change it later. I want to do a webapp that displays a cartoon map of the forest in Colorado. Users can choose values for each feature, and a tree of a particular type will pop up on the map along with some fun facts about this species. I need webapp development skills and user interface design skills for this application.

[1] “Tackling Climate Change with Machine Learning,” Accessed on: Feb 1, 2021. [Online]. Available: arXiv:1906.05433

[2] UCI Machine Learning, “Forest Cover Type Dataset,” in *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/uciml/forest-cover-type-dataset, Accessed on: Feb. 2, 2021.

[3] Dua, D. and Graff, C., “Covertype Data Set,” in *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science, 2019. Available: http://archive.ics.uci.edu/ml, Accessed on: Feb. 2, 2021

[4] D. Lakshmanan, *How, When, and Why Should You Normalize / Standardize / Rescale Your Data?*, towards AI. Accessed on: Feb 7, 2021. [Online]. Available: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff#:~:text=Normalization%20is%20useful%20when%20your,Gaussian%20(bell%20curve)%20distribution.

[5] P. Shivaprasad, *Understanding Confusion Matrix, Precision-Recall, and F1-Score*, towards data science. Accessed on: Feb 7, 2021. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-precision-recall-and-f1-score-8061c9270011