Final Project Critique: Rohit Yadav’s Climate Analysis and Prediction

By Luke Sturgeon

**Intro**

Rohit used climate data from Kaggle to try and predict weather patterns in India. This was a big data set and was thoroughly analyzed using regression, clustering, and association techniques to generalize how the variables are related.

**Best Point**

Rohit’s best point, in my opinion, was pointing out how much better his classification models performed when he used the original data rather than the principle components he found after his PCA, despite these two components explaining 98.44% of the variation in the data.

**Weakest Point**

I would have liked to see his explanation as to why the 1.66% of variance left out of these two principle components cause such a large change in his models’ predictive power.

**Dataset Size**

His dataset was very large. In his paper, his EDA shows the dataset to be 100990 x 14, or a little over 1.4 million cells before any data reduction techniques. Even after removing superfluous data, his data set remained well over the required dataset size for this project.

**Relevancy**

This project is very relevant to what we learned in class. Machine learning is widely used in weather prediction models and can greatly improve the predictive capabilities within this field; weather prediction is rather infamous for poor predictability.

**Difficulty**

This dataset had a generous mix of data types, making working with it challenging. It contained categorical, numeric (both integer and continuous), as well as text data. It’s a rich dataset but that means a lot of pre-processing must be done for each analysis type.

**Interestingness**

Because of the variety of data types within the dataset and the relevancy to everyday life, this project piqued my interest. Not only does it have a real-world impact, something that may be used today, but managing the data to be used in the different processes was interesting to see as well.

**Clarity**

Rohit used a lot of EDA and using two principle components for parts of his analysis gave us a lot of eye candy in colorful charts and graphs the helped explain the points he was making. Very clear for how convoluted and strangely behaving this dataset is.

**Originality**

I’m not familiar with any weather pattern models that have included text data as well as numeric and integer. This addition gives an original perspective to modelling this data.

**Creativity**

His use of graphs, especially when it came to his association rules, was very creative. The colors and types of graphs he used made the project more interesting and easy to understand because of them.