**Navigating the Flood Vulnerability: Data-Driven Analysis for Assam’s Revenue Circles**

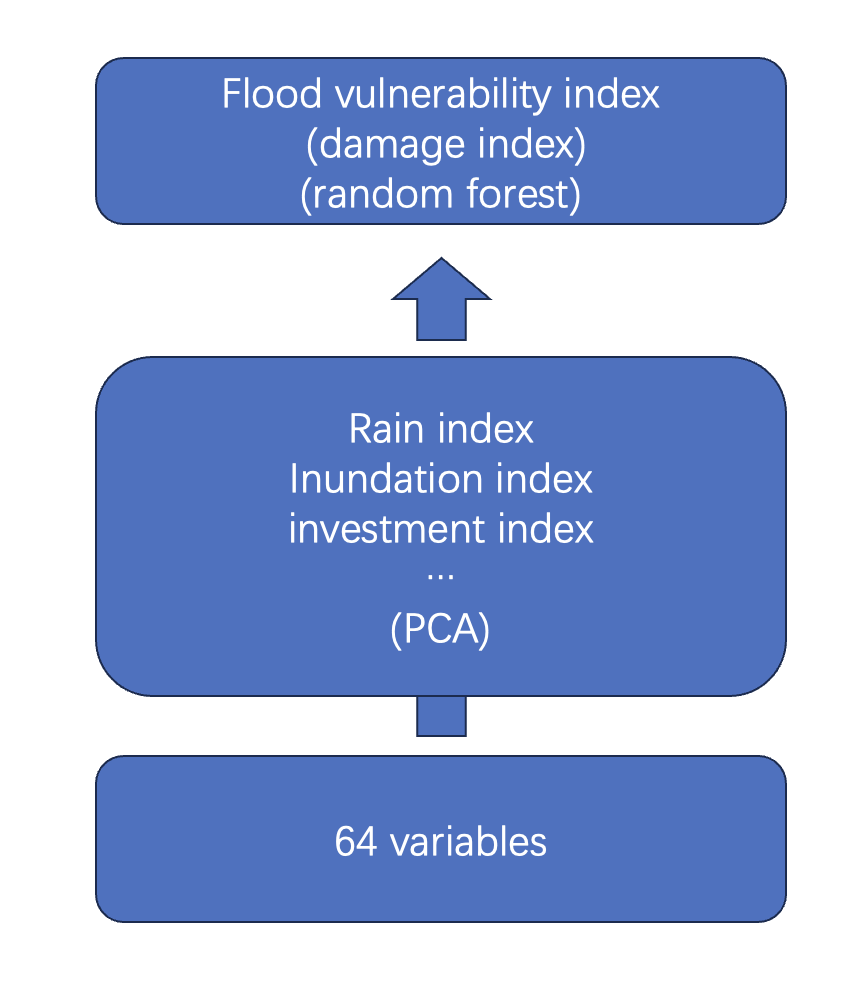
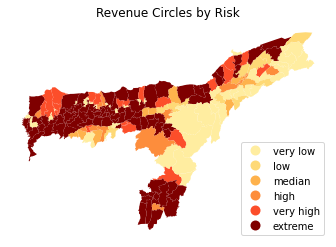
Every year, the state of Assam in India faces a serious problem with floods. Nearly 40% of the state is at risk, leading to significant damage to people's lives, their livelihoods, and the infrastructure. This constant threat of flooding is a major concern for the people of Assam. In the past, the funds allocated for disaster relief in the state have fallen short of what's needed, partly due to various factors influencing the decision-making process. Consequently, there is an urgent need for a better way to allocate these funds efficiently and enhance flood preparedness across Assam. We have developed two classification models to categorize the 180 revenue circles in Assam into six groups based on their vulnerability to floods. These models aim to assess the flood vulnerability of each revenue circle and provide recommendations for allocating disaster relief funds more efficiently. The goal is to enhance overall flood preparedness and resilience throughout the state of Assam, ultimately ensuring better protection for its residents and resources.

We have a dataset that contains information about floods in different areas from May 2021 to August 2023. This data includes 64 different aspects related to floods, and we've organized them into six categories to better understand the vulnerability, risk, resilience, and readiness in each area:

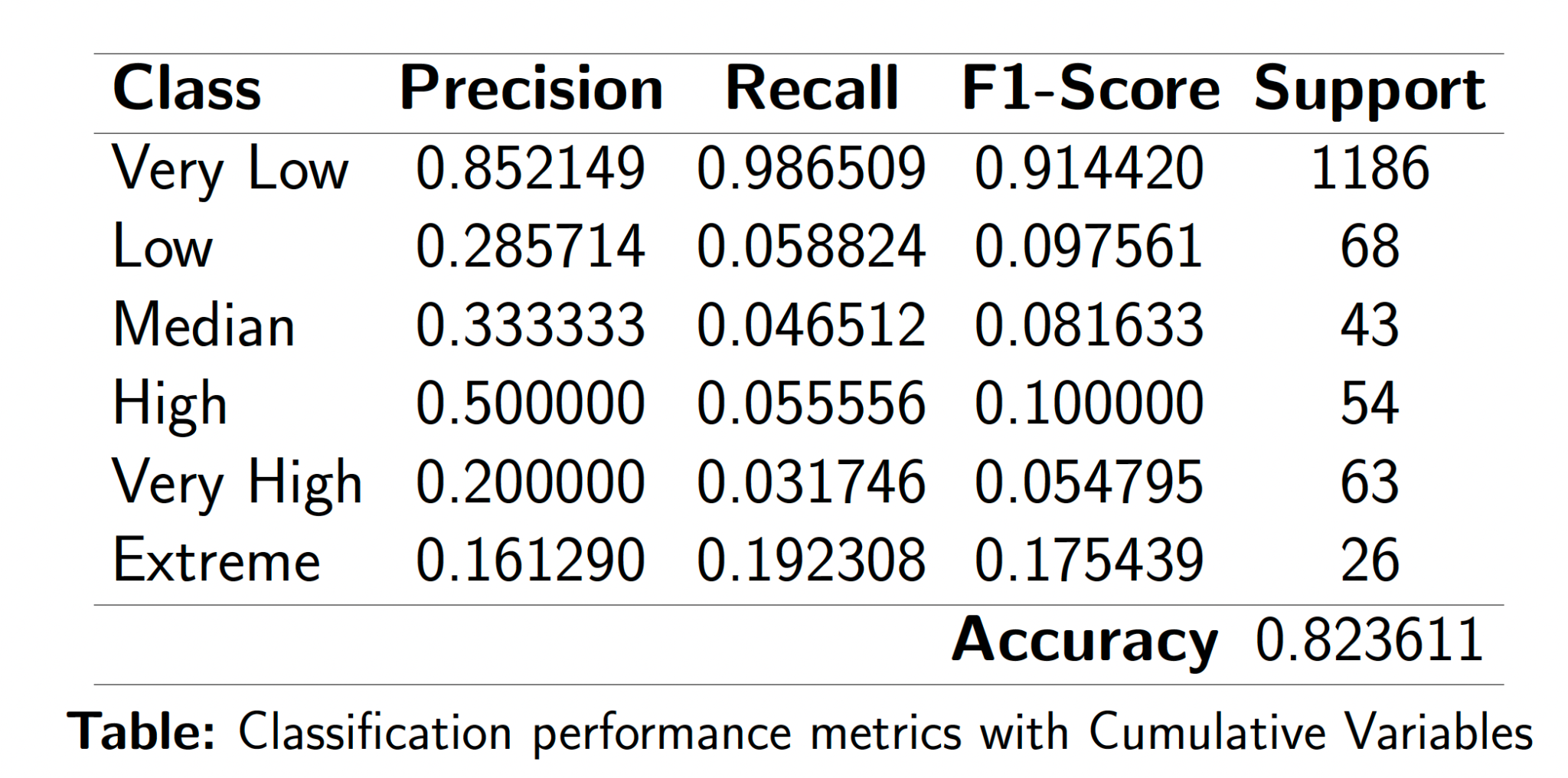
1. Flood Proneness Variables: These include factors like elevation, slope, and drainage density that help us assess how susceptible an area is to flooding.
2. Socio-economic Vulnerability Variables: This category looks at aspects like the age distribution in the population, particularly focusing on older and younger residents who may be more vulnerable during floods.
3. Demographic Variables: These variables consider factors like the total population and the ratio of males to females in the area.
4. Environmental Vulnerability Variables: This category examines the presence of critical infrastructure such as roads, schools, and hospitals, which can be affected by floods and impact the community's response and recovery.
5. Government Responses Variables: We look at indicators like the number of relief camps and the financial assistance provided by the government in response to floods.
6. Damages and Losses Variables: This category includes data on the human toll of floods, such as lives lost, as well as damage to roads and bridges.

By organizing these flood-related variables into groups, we are able to extract valuable insights from each group using a technique called Principal Component Analysis. This helps us gather crucial information from various aspects of data related to Assam state, enabling us to make smarter decisions to improve our readiness and resilience in dealing with floods. We've identified eleven key factors related to floods, each representing a different aspect: damage, government investment, inundation, rainfall, river behavior, population distribution, road conditions, terrain features, land characteristics, drainage systems, and electricity availability.

In the next phase of our work, we'll use these factors as the foundation for creating classification models. To kick things off, we'll begin our flood damage analysis by applying the Random Forest method to the results of above extracted indices. This Random Forest model is designed to predict the potential damage that each revenue circle might experience if a flood were to occur.

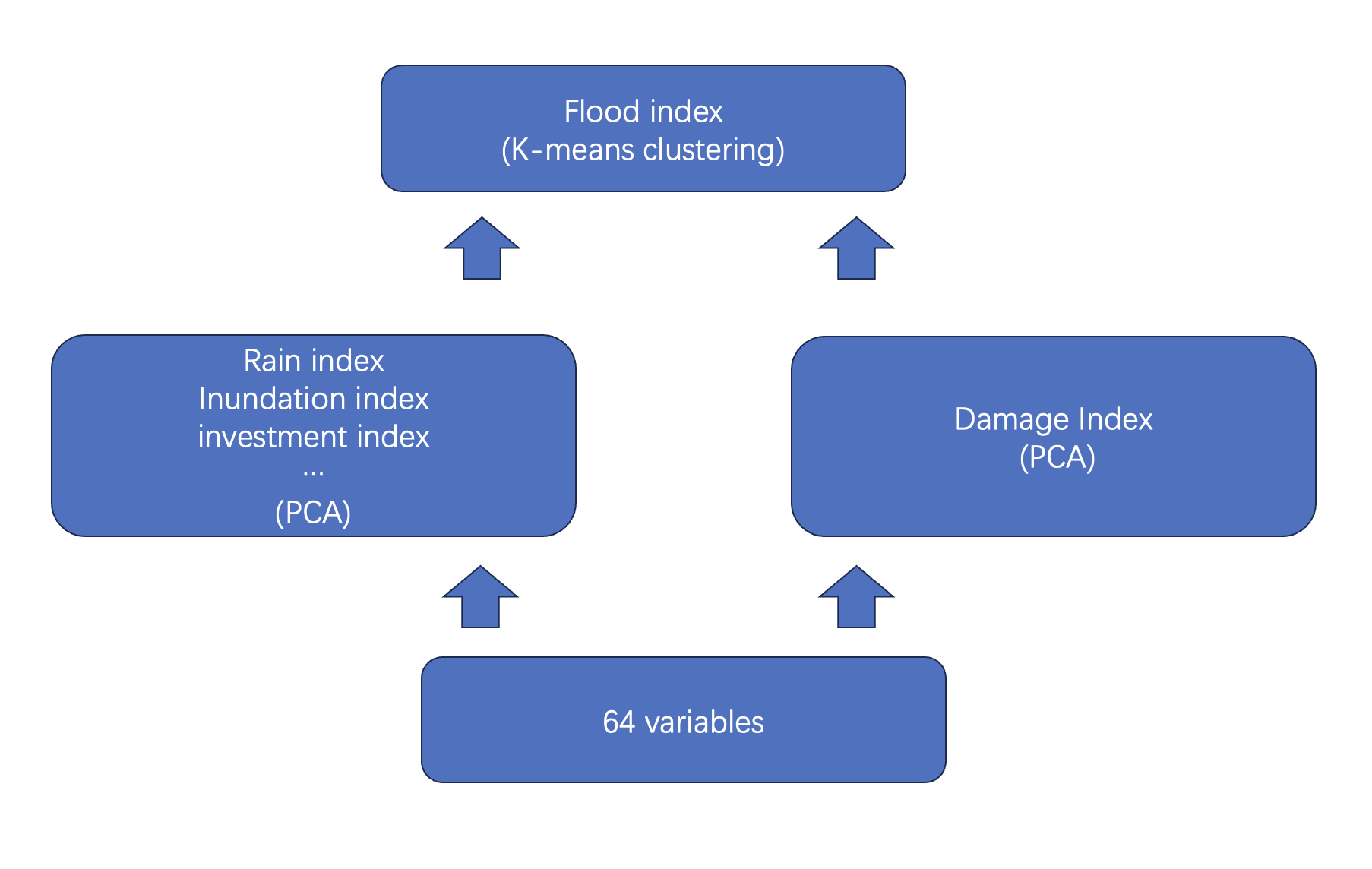


Random forest prediction at 2022.06 Random forest model structure



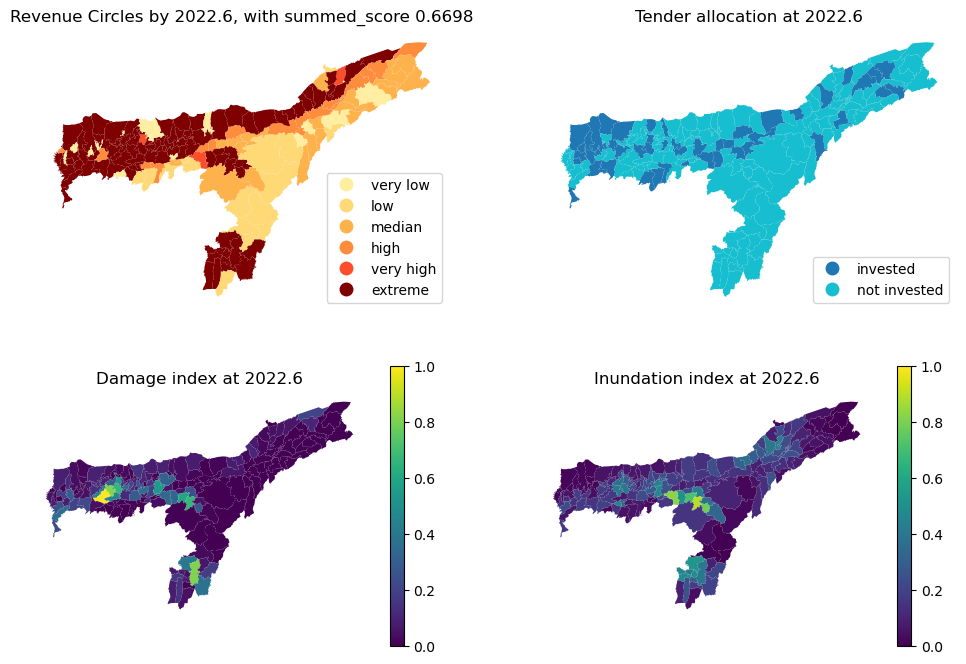
The random forest model seems to achieve a relatively high accuracy rate, approximately 82%. However, it's important to note that this accuracy is somewhat skewed due to the model's excellent performance on the very low class, which has a significantly larger number of instances.

We also used a technique called K-means clustering to create a more general flood index, using information from the Principal Component Analysis (PCA) indices. The key distinction between the K-means clustering and Random Forest methods is how they use the PCA results and the damage index: In the case of Random Forest, it's focused on predicting the flood damage index by using the PCA results, excluding the damage index itself. On the other hand, K-means clustering considers all the PCA indices, including the damage index, to create a more general flood index that captures various aspects related to flooding.



K-means model structure

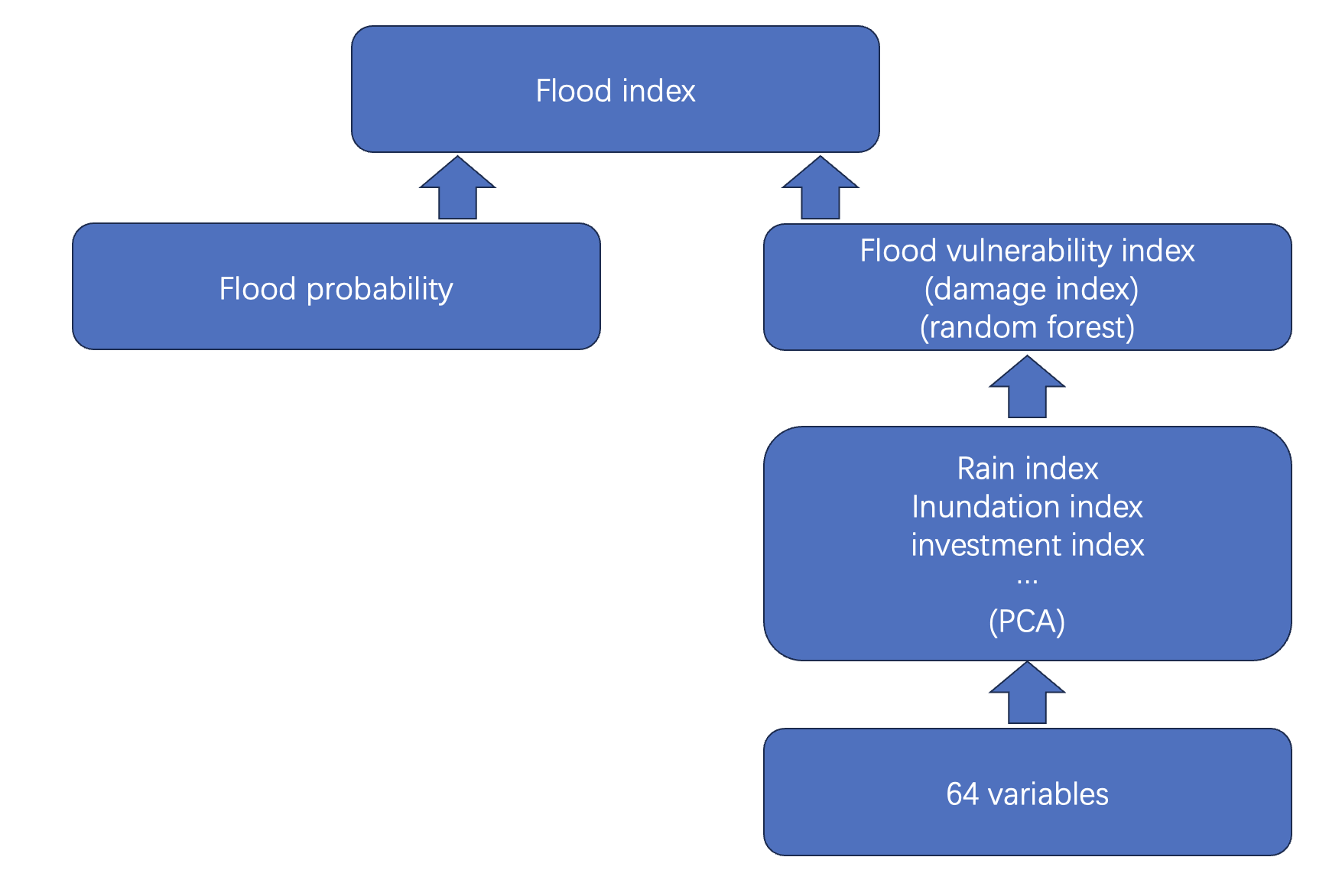
Instead of ranking entities, K-means clustering groups them based on their similarities. To give this clustering a ranking aspect, we combined important indices like the government investment index, damage index, inundation index, and rain index into a single score. We then used this combined score to rank the K-means groups. This classification approach is valid because our flood index accurately reflects the level of damage, inundation, and rainfall in specific areas, and it helps identify inconsistencies in previous resource allocation.



K-means clustering at 2022.06

It's worth noting that we used an unsupervised K-means algorithm to create this flood index, and while it's promising, we plan to conduct statistical verification in the future. For now, we're evaluating our results through visual comparisons.

In conclusion, the frequent and devastating floods in Assam, India, call for a shift away from relief allocations influenced by politics and towards a more data-driven approach. Our Random Forest classifier has shown high accuracy and provided a reasonably good flood vulnerability map. However, it's important to note that its accuracy is somewhat influenced by its performance on the dominant "very low" class. The K-means results have effectively mirrored the real-world impact of floods, including damage, inundation, and rainfall, while also revealing disparities in past resource allocations.



Potential future model structure

Nevertheless, our next crucial step is to establish a method for validating these results to ensure their reliability. Looking ahead, in our next phase of work, we can develop a flood prediction model that estimates the likelihood of a flood occurring. We can then combine this prediction with the flood vulnerability information from the Random Forest model to create a more comprehensive flood index. This new index would provide a broader perspective on the flood risk. We can then use this enhanced flood index to cross-verify the flood index generated from the K-means clustering model, ensuring the accuracy and reliability of our assessments.

We hope that our research offers valuable insights, and we hope that other researchers can use these datasets and models for a wide range of purposes in a more comprehensive way. Our goal is to contribute to the well-being of the people of Assam and beyond by making this data available for various applications and initiatives.

This article is part of CivicDataLab and New York University’s Capstone project.