- DIRECT Project -

EASE (Electricity Analysis Suggestion Ensemble)

Technology Review

github.com/danielfather7/EASE-Project

Team 6

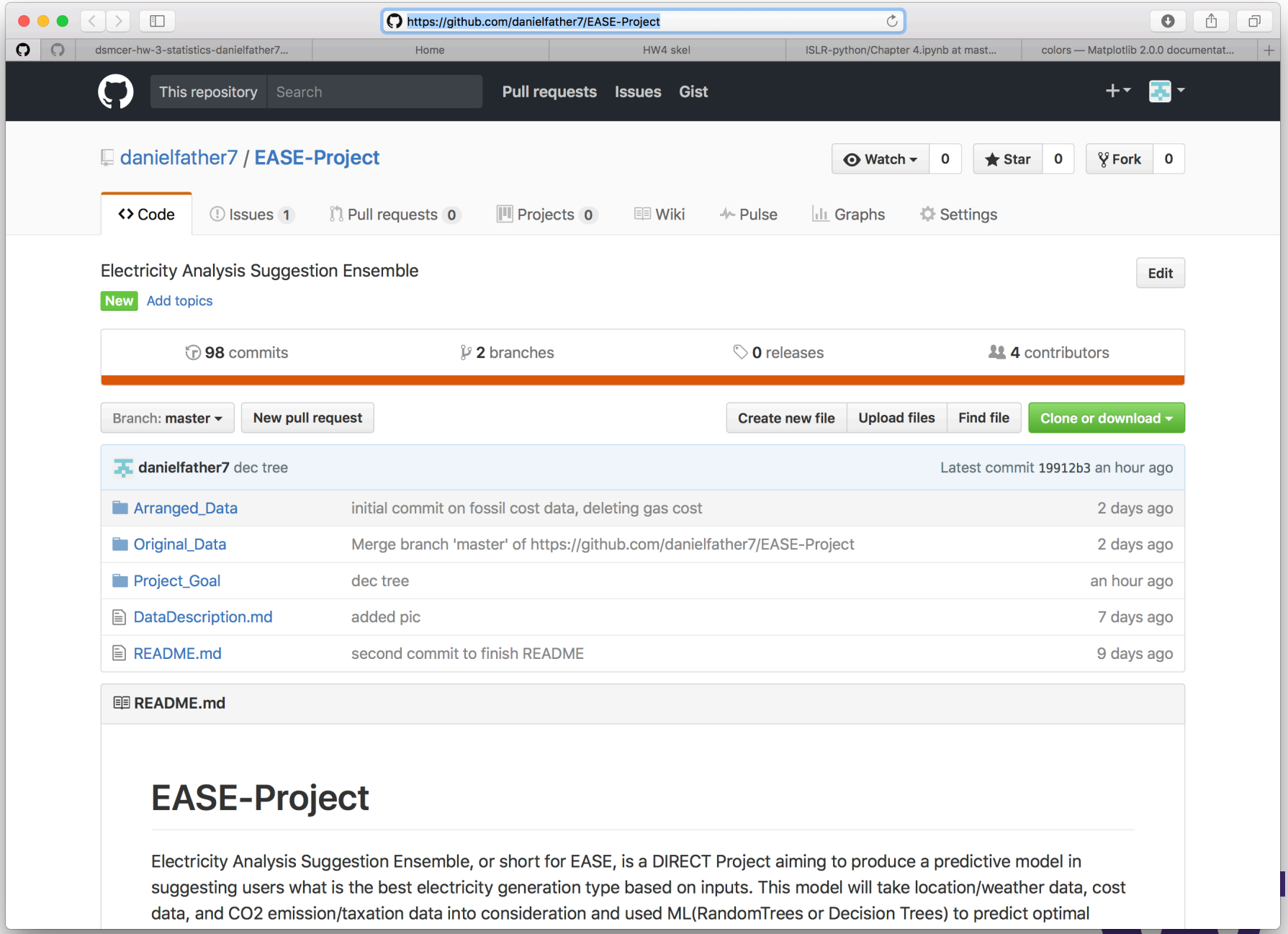
Ivan Cui

Daniel Pan

Yongquan Xi

Jiayuan Guo

UNIVERSITY of WASHINGTON

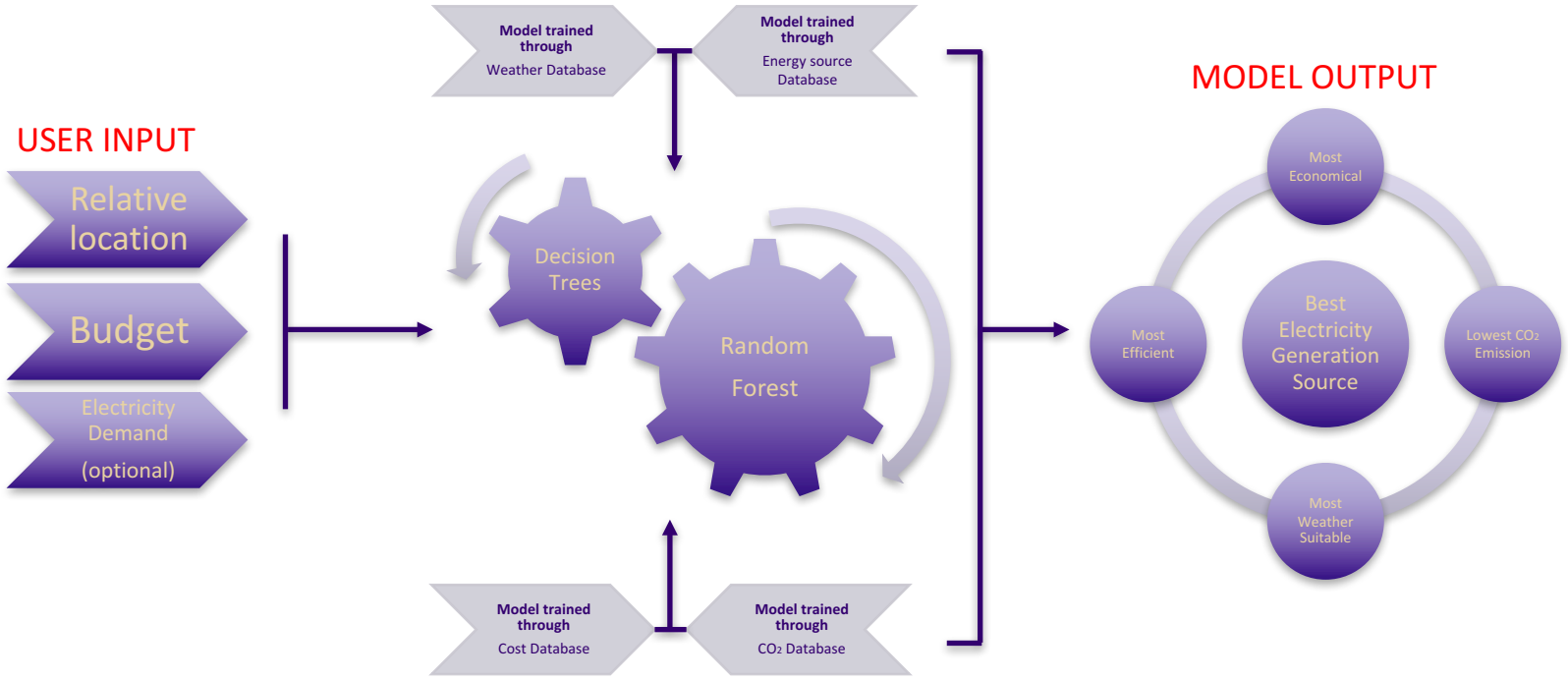

https://github.com/danielfather7/EASE-Project
dsmcer-hw-3-statistics-danielfather7...
Home
HW4 skel
ISLR-python/Chapter 4.ipynb at mast...
colors — Matplotlib 2.0.0 documentat...
This repository
Search
Pull requests
Issues
Gist
danielfather7 / EASE-Project
Watch 0
Star 0
Fork 0
Code
Issues 1
Pull requests 0
Projects 0
Wiki
Pulse
Graphs
Settings
Electricity Analysis Suggestion Ensemble
Edit
New
Add topics
98 commits
2 branches
0 releases
4 contributors
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
danielfather7 dec tree
Latest commit 19912b3 an hour ago
Arranged_Data
initial commit on fossil cost data, deleting gas cost
2 days ago
Original_Data
Merge branch 'master' of https://github.com/danielfather7/EASE-Project
2 days ago
Project_Goal
dec tree
an hour ago
DataDescription.md
added pic
7 days ago
README.md
second commit to finish README
9 days ago
README.md
EASE-Project
Electricity Analysis Suggestion Ensemble, or short for EASE, is a DIRECT Project aiming to produce a predictive model in suggesting users what is the best electricity generation type based on inputs. This model will take location/weather data, cost data, and CO2 emission/taxation data into consideration and used ML(RandomTrees or Decision Trees) to predict optimal

Background-Project

- Electricity is one of the major energy which can be generated through different resources:
  - Conventional Resources: Coal, Natural gas, Petroleum, etc. Limitation: High CO_2 emission & accordingly high CO_2 taxation
  - Clean Energy Resources: Solar, Wind, Hydro, Biofuel etc. Limitation: Technology & Distributive location of particular resources
- Electricity generation source model to provide suggestion on the electricity generation type:
  - Factor 1: Weather (including Temperature, Precipitation, Wind speed)
  - Factor 2: Financial Cost (Cost per watt by different resources)
  - Factor 3: CO_2 Tax per year depending on different states

Project Objective

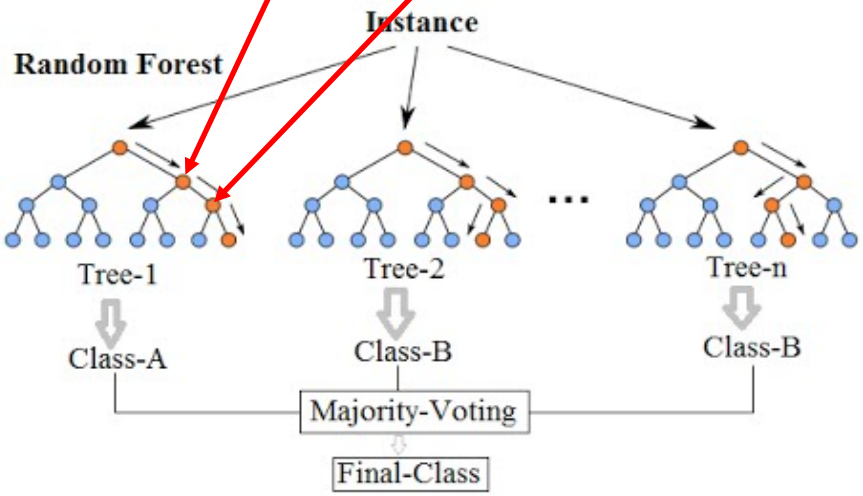
> Develop a predicative model using ML to provide users suggestions on the best electricity generation source type.



Package & Algorithm

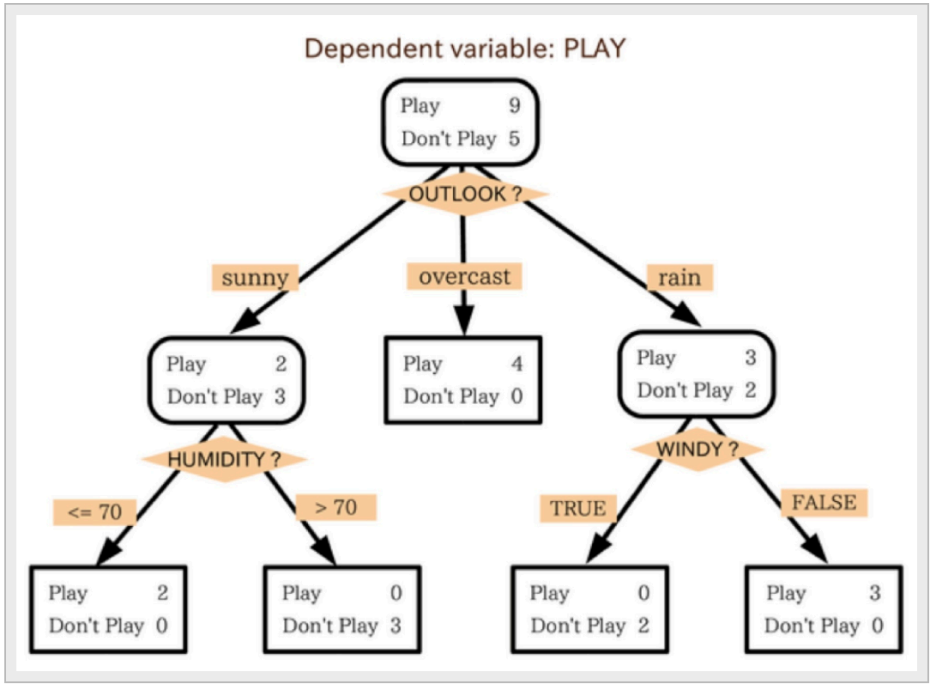
- Package: Scikit-learn
  - Open-source machine learning modules
  - Built-in dataset and various algorithm for classification & regression
  - Useful for dataset loading, transformation, features selection, etc.
- Learning Algorithm: Random Forest
  - One of the most popular and accurate learning algorithms available
  - Combine Bagging method and random decision trees, outputs the class that is the majority voting by individual trees

Latitude	Longitude	AvgTemp	TotalMonthlyPrecip
51.883	-176.650	51.6	0.22
30.688	-88.246	81.7	4.95



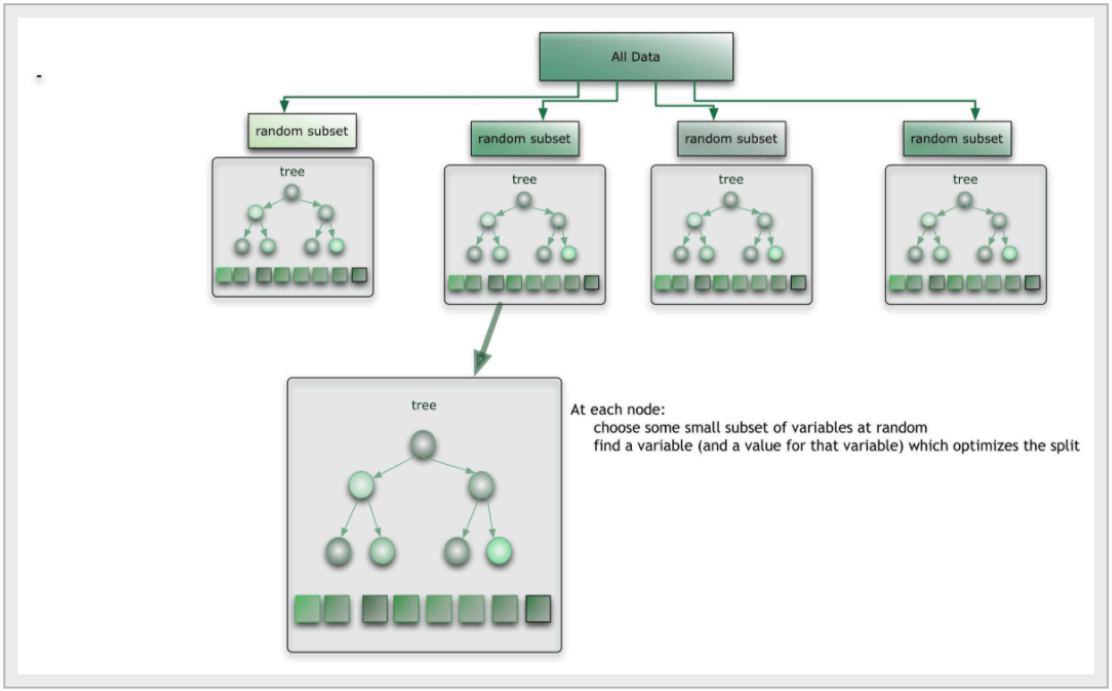
How it works

- Decision trees are individual learners that are combined. They are one of the most popular learning methods commonly used for data exploration.



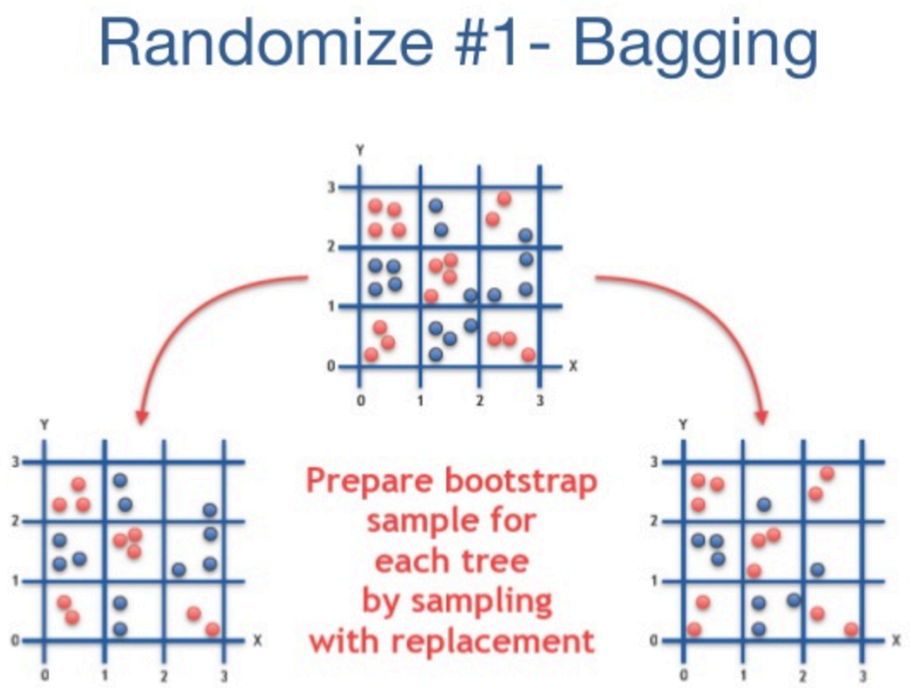
How it works

- RandomForests consists of an ensemble of classification decision trees, and outputting the mode of the classes of the individual trees.
  - Corrects Decision Tree's overfitting and inaccuracy by voting.



How it works

- Tree-bagging is an algorithm that selects a random subset of the features at each candidate split in the learning process, a build-in feature to de-correlate individual trees.
  - Decrease variance of the model without increasing bias.



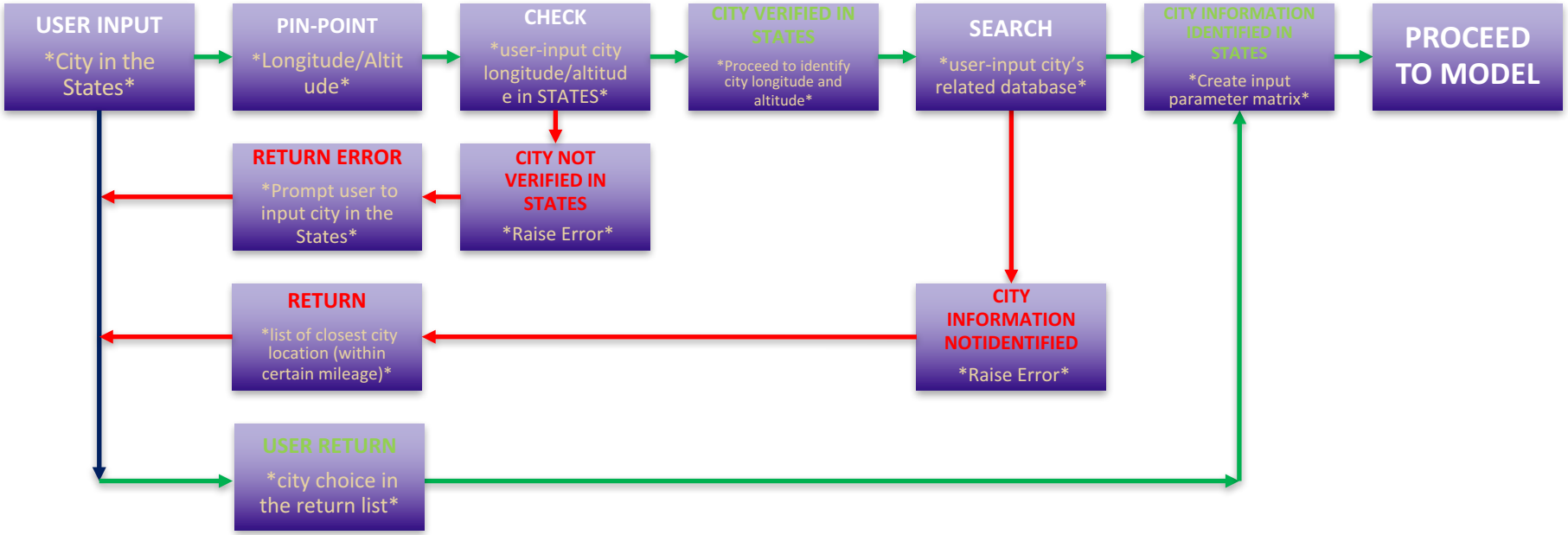
Appeal

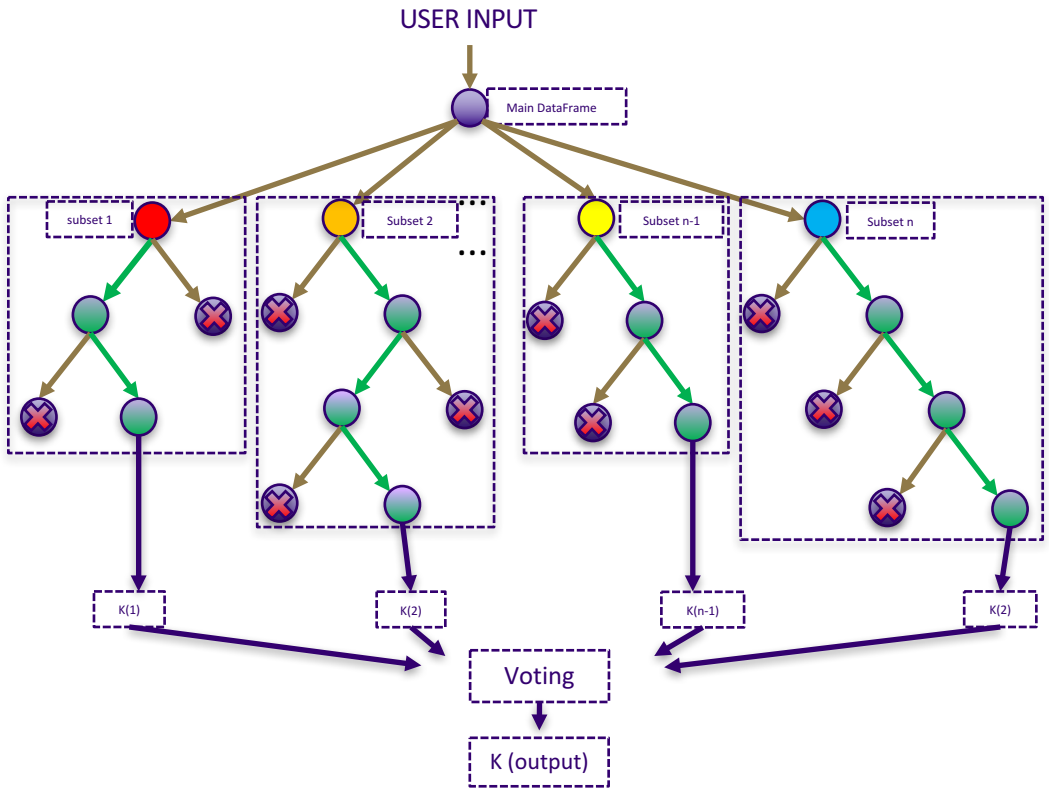
- Easy to learn, fast to build.
  - Project limitation.
- One of the most accurate learning algorithms.
- Run efficiently on large databases, and able to maintain accuracy even when a large proportion of the data are missing.
  - Our weather data （262477 rows × 8 columns)
- Give estimates of what variables are important in the classification.
- Generate an internal unbiased estimate of the generalization error as the forest building progresses.
  - Out of bootstrap (OOB) samples.
  - Estimated test error is very accurate in practice, with reasonable N.
  - Another validation set or cross-validation is not required, speeds up training.

Drawbacks (RandomForests)

- Fast to train data, but slow to predict
- More trees are required to generate a higher accuracy, which provides poor run-time performance. It becomes an issue when faster algorithm is preferred
- A predictive modelling tool instead of a descriptive tool – hard to interpret the information extracted from the trained forest
- The situation of overfitting when you have large number of categorical variables with different levels - Larger the tree, more overfitting for training data

User Case Example




USER INPUT
Main DataFrame
subset 1
Subset 2
Subset n-1
Subset n
K(1)
K(2)
K(n-1)
K(2)
Voting
K (output)