VANDERBILT ISIS SUMMER INTERN -- DATA IMPUTATION FOR WEATHER

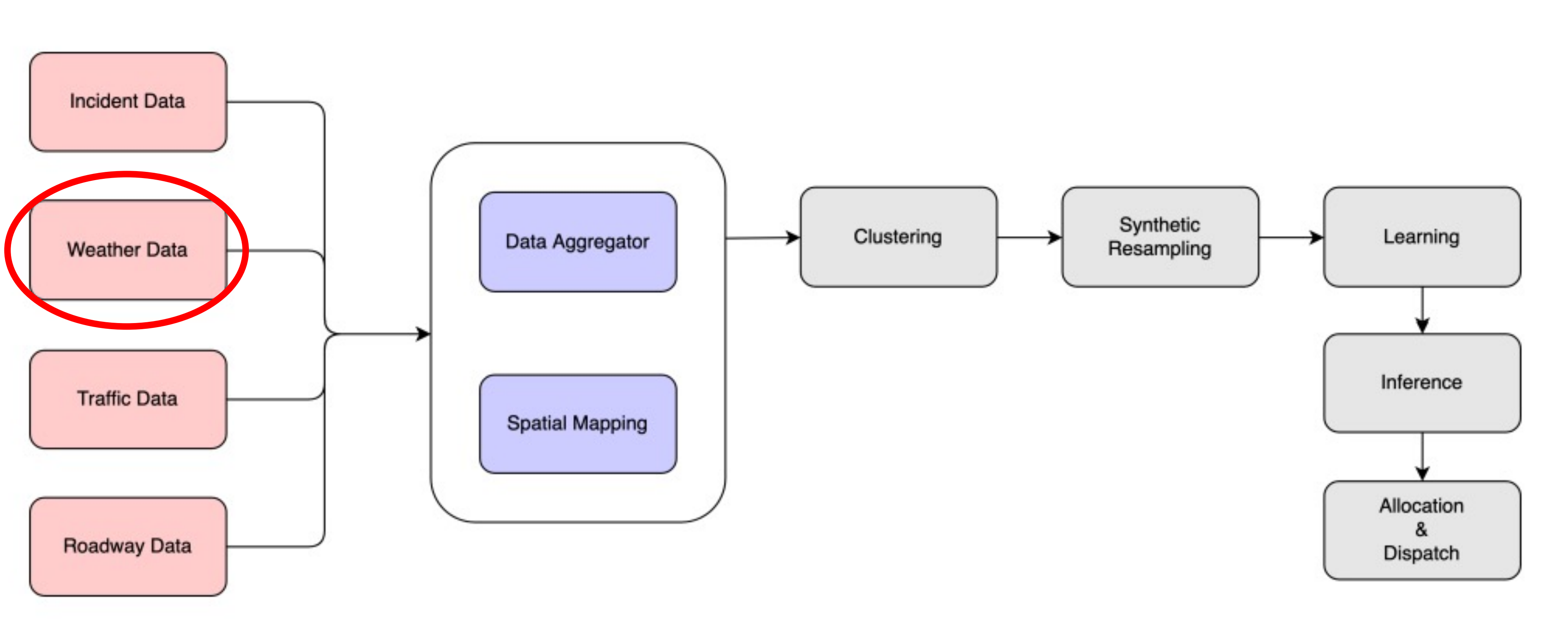
-- Vera Yang

Mentors: Abhishek Dubey

Sayyed Mohsen Vazirizade

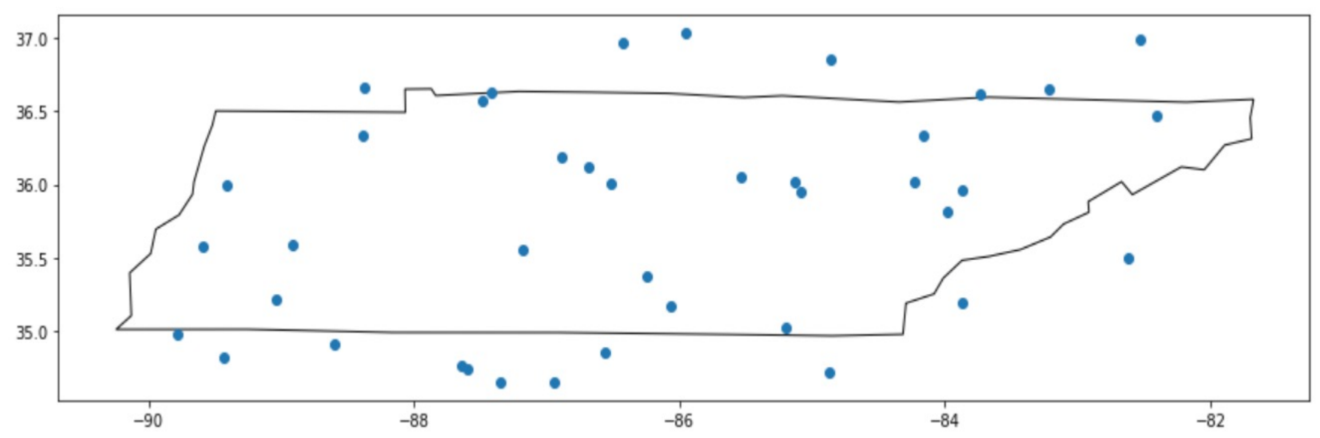
VANDERBILT UNIVERSITY

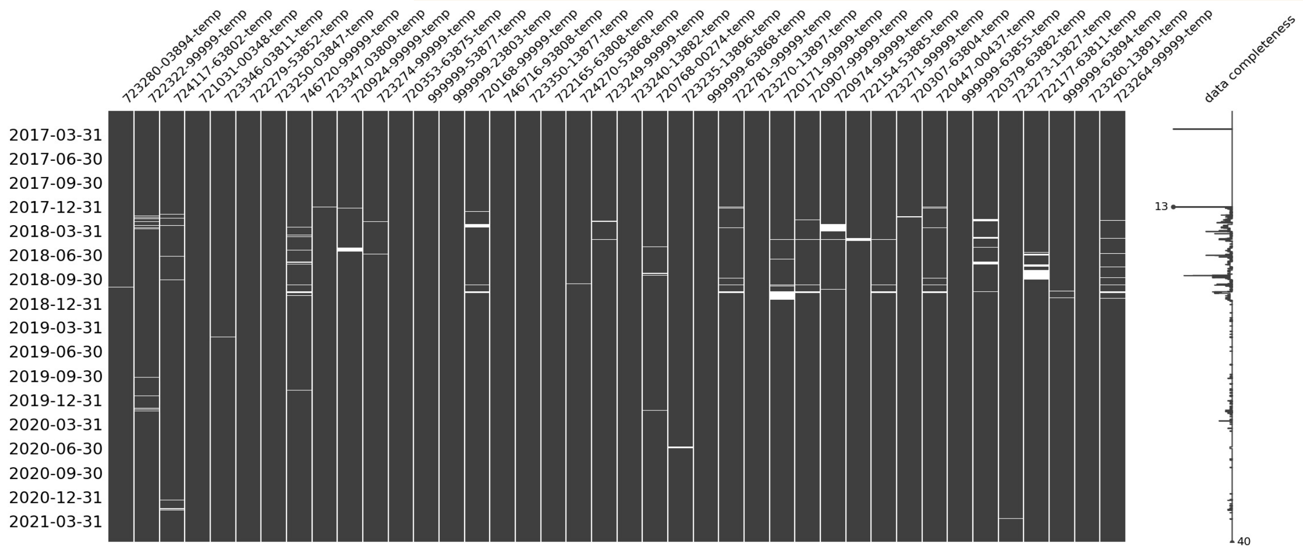
Incident Prediction



--Data preparation (collection, cleaning, imputation, feature selection, analyzing) becomes the very first step.

Weather Data



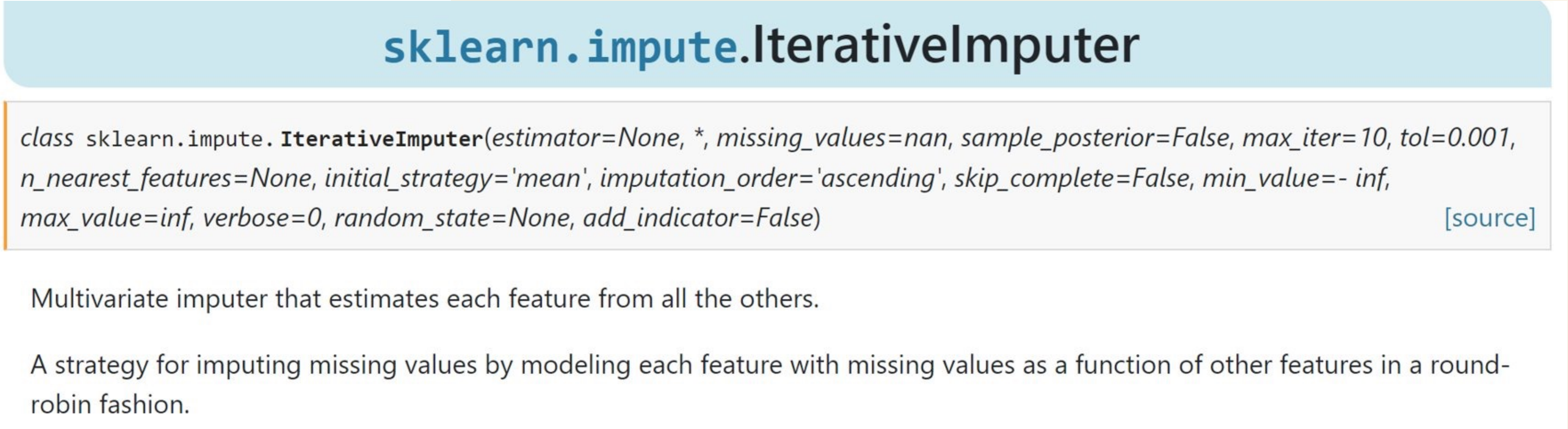


- Weather data (temperature, precipitation, visibility, wind speed…) for 40 weather stations in or around Tennessee.

- Missing values in the temperature feature, plotted with missingno.

VANDERBILT UNIVERSITY

Multivariate Imputation by Chained Equation (MICE)



sklearn.impute.IterativeImputer

class `sklearn.impute.`**`IterativeImputer`**(*estimator=None, *, missing_values=nan, sample_posterior=False, max_iter=10, tol=0.001, n_nearest_features=None, initial_strategy='mean', imputation_order='ascending', skip_complete=False, min_value=- inf, max_value=inf, verbose=0, random_state=None, add_indicator=False*) [source]

Multivariate imputer that estimates each feature from all the others.

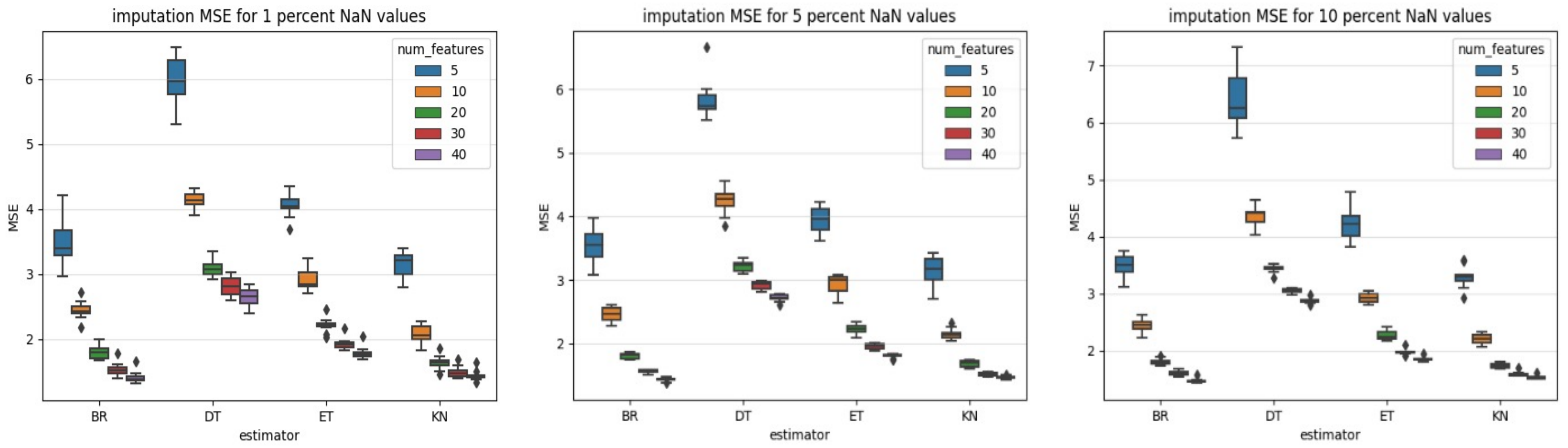
A strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion.

Hyperparameter search:

- Estimator: BayesianRidge(BR), DecisionTreeRegressor(DT), Extra TreeRegressor(ET), KneighborsRegressor(KN)
- n_nearest_features: 5, 10, 20, 30 ,40

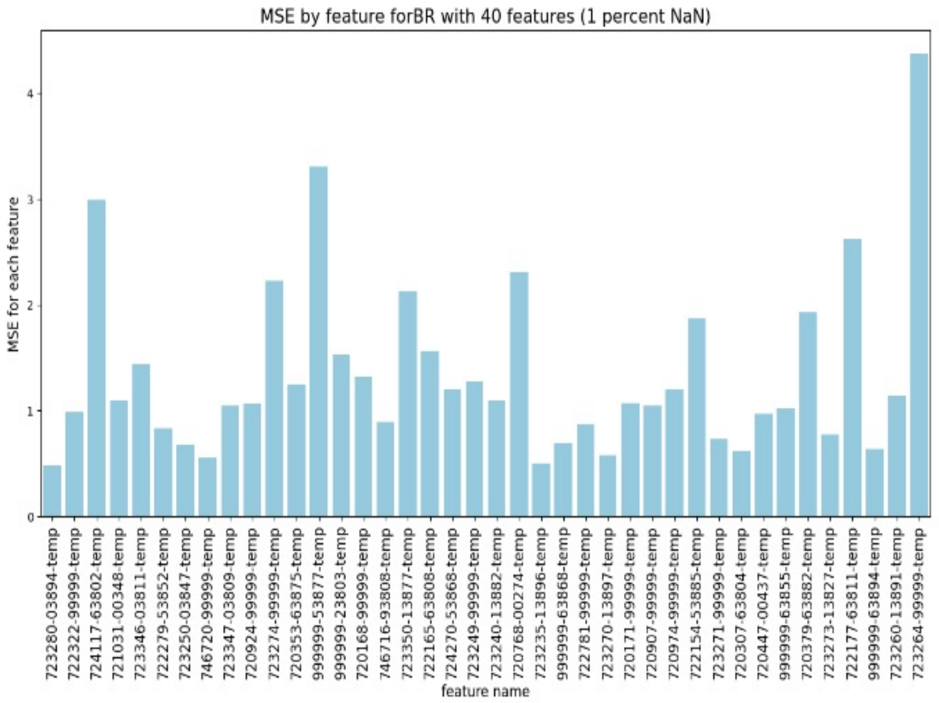
--The goal is to find the best combination of these hyperparameters based on the smallest MSE.

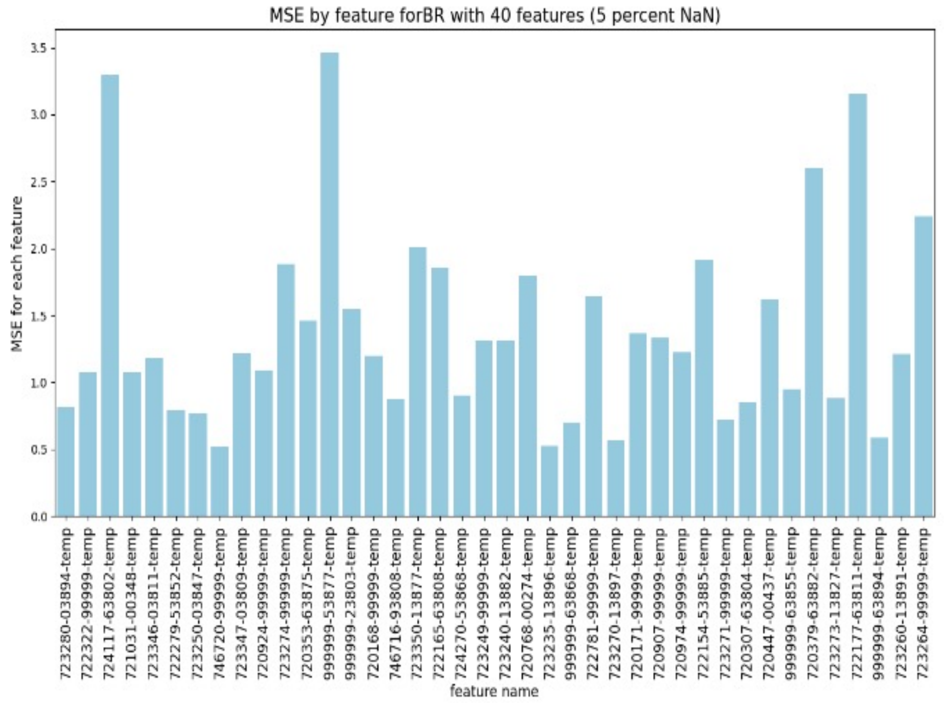
Results

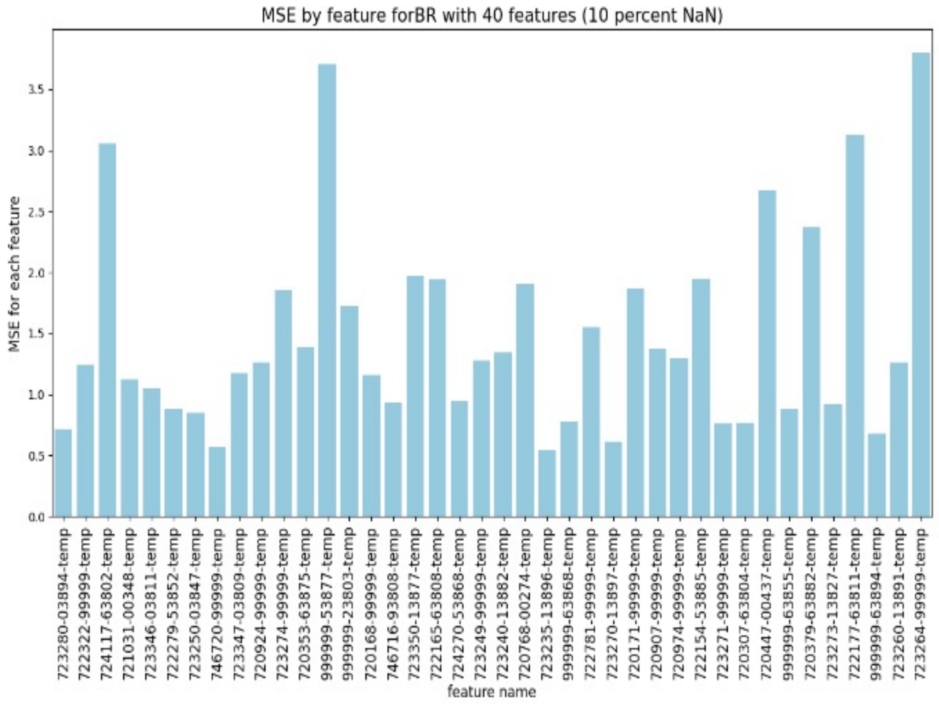


- Good models are first chosen by looking at the overall MSE. Here, BR, ET, and KN perform better than DT; also, MSE decreases as we use more features.
- BR with 40 features and KN with 40 features are the 2 best models. Since BR is much faster than KN, we choose BR with 40 features.

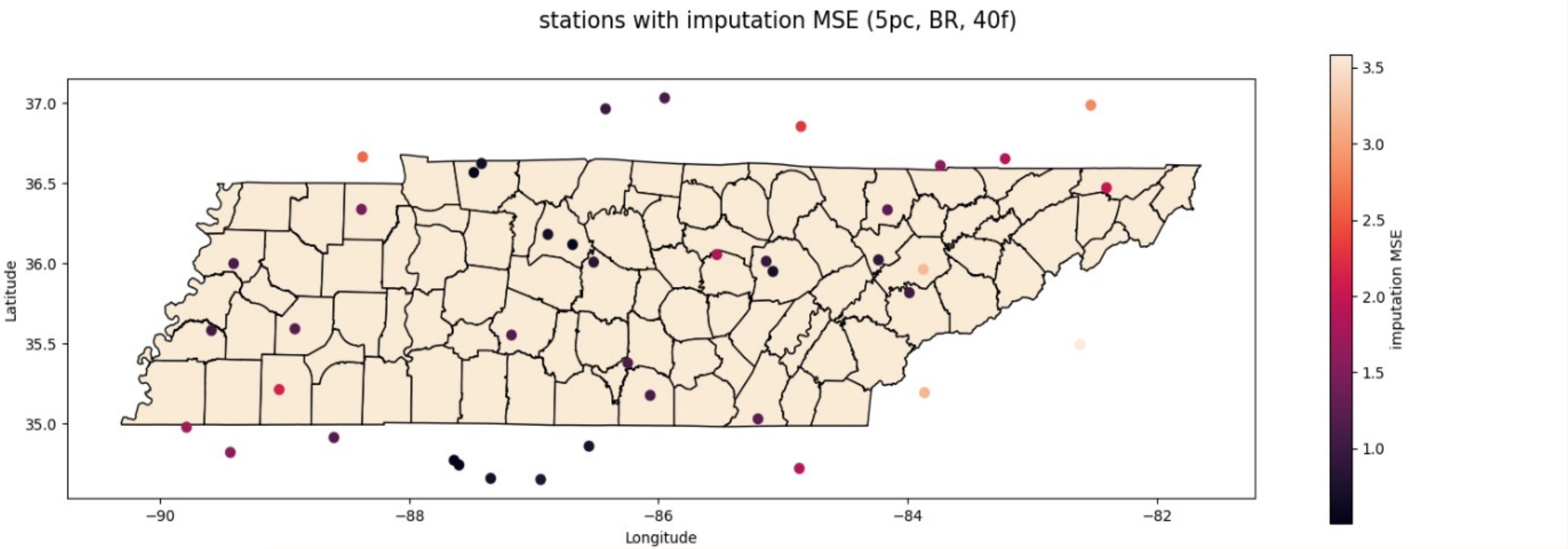
VANDERBILT
UNIVERSITY







- Ideally, we want the station-wise MSE to be a uniform distribution.



- In general, the NA values in stations having one or more close neighbors are imputated more accurately.

Future Improvement on Imputing Precipitation Data

- Problem: 80% percent of values are zeros; imbalanced dataset
- Solution: Binary classification with performance evaluated by F-measure

Thoughts and Reflections

- Readings helped
- Coding and Pipeline Design
- Help from the team

VANDERBILT
UNIVERSITY

THANK YOU

VANDERBILT
UNIVERSITY