Climate-Based Forecasting of Dengue Epidemic Months: A Case Study of Bangladesh

THE ERDŐS INSTITUTE
Revolutionary Collaborations in Academia and Industry

Haridas Kumar Das & Abdullah Al Helal

A Data Science Bootcamp project

Erdős Institute's May-Summer 2024 Cohort!

Overview

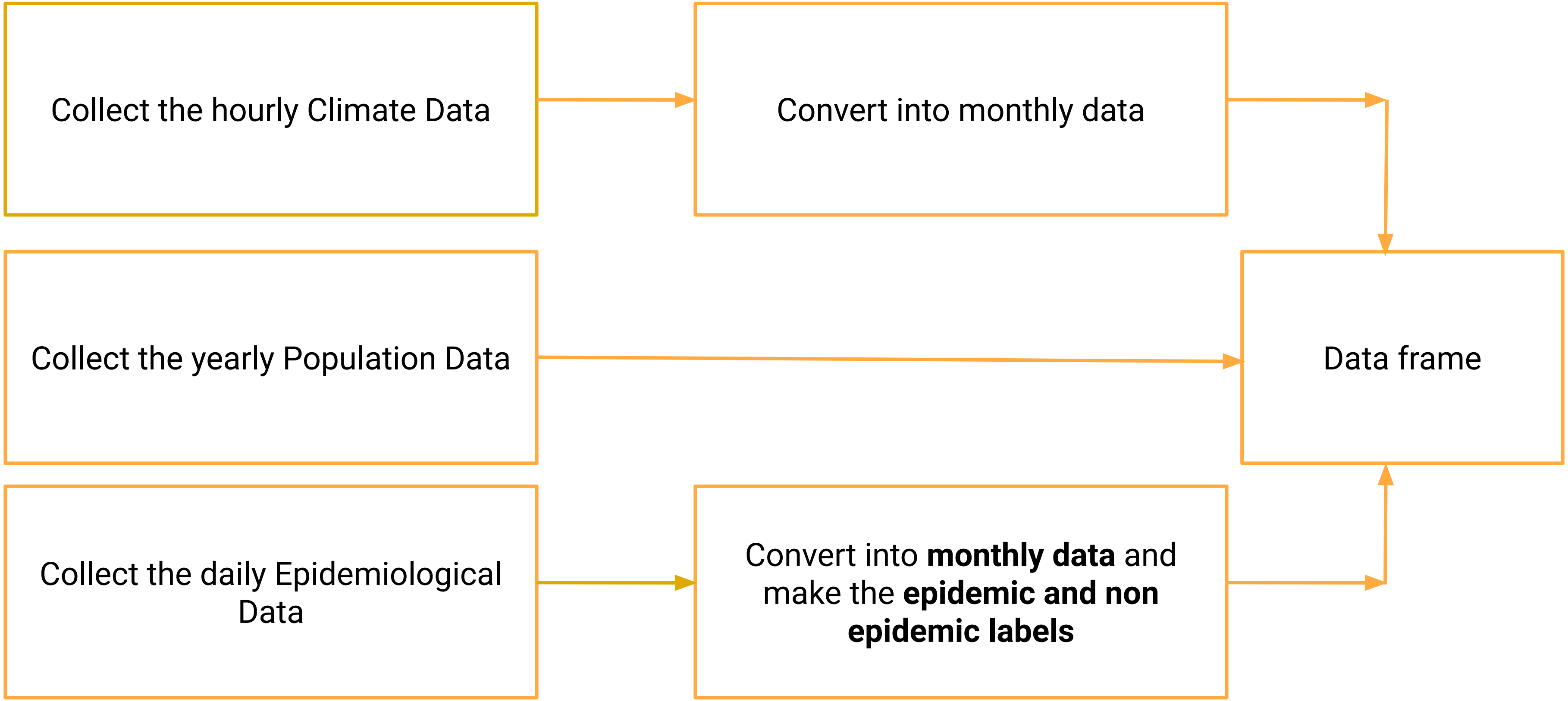
Question: How can we develop machine learning algorithms to analyze climate and epidemiological data in order to forecast epidemic months of diseases, for example, dengue?

Goal: We develop machine learning algorithms to analyze climate and epidemiological data in order to forecast dengue epidemic months, focusing on the analysis of Bangladesh.

Use: Policymakers can use our model to analyze or predict epidemic diseases in the decision-making process.

Data: Climate Data, Population Data, and Epidemiological Data

Data preprocessing



Collect the hourly Climate Data

Convert into monthly data

Collect the yearly Population Data

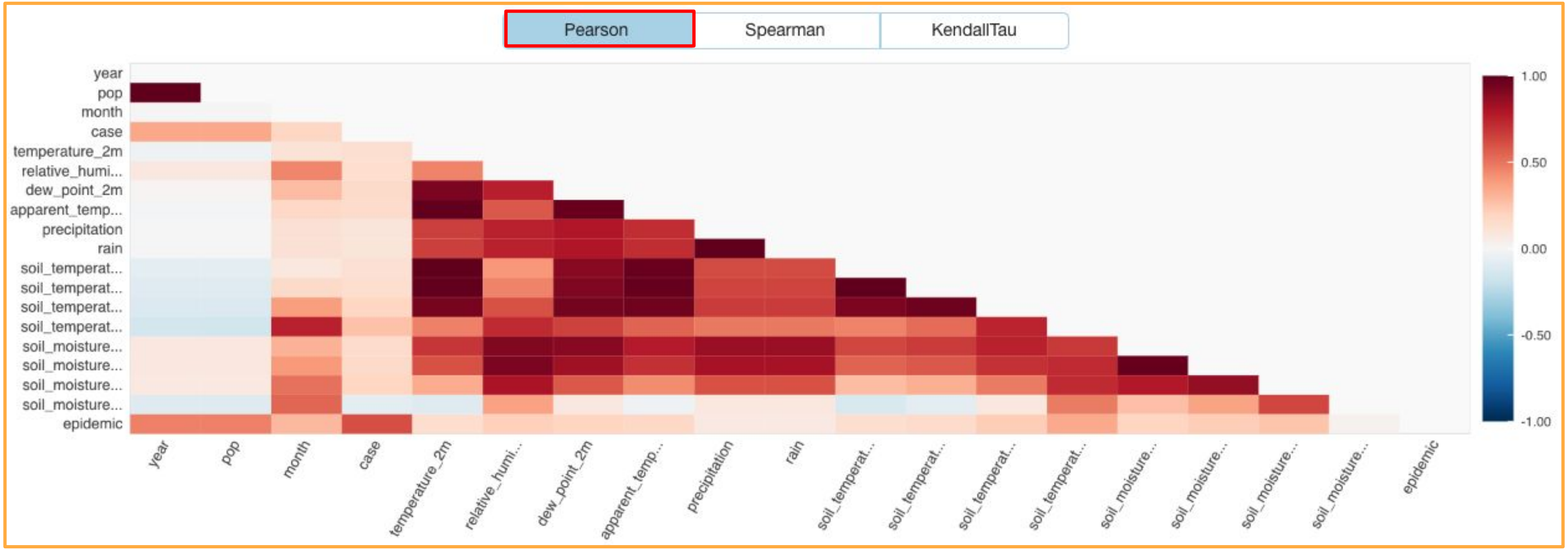
Data frame

Collect the daily Epidemiological Data

Convert into **monthly data** and make the **epidemic and non epidemic labels**

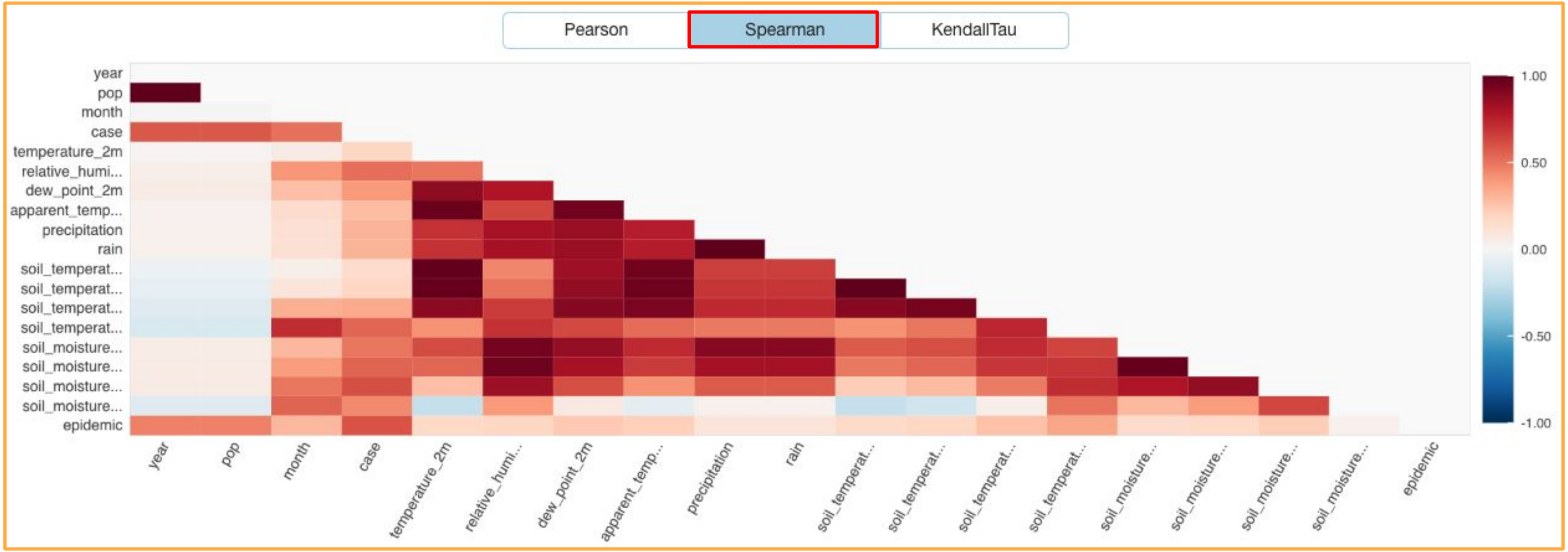
Exploratory Data Analysis

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry



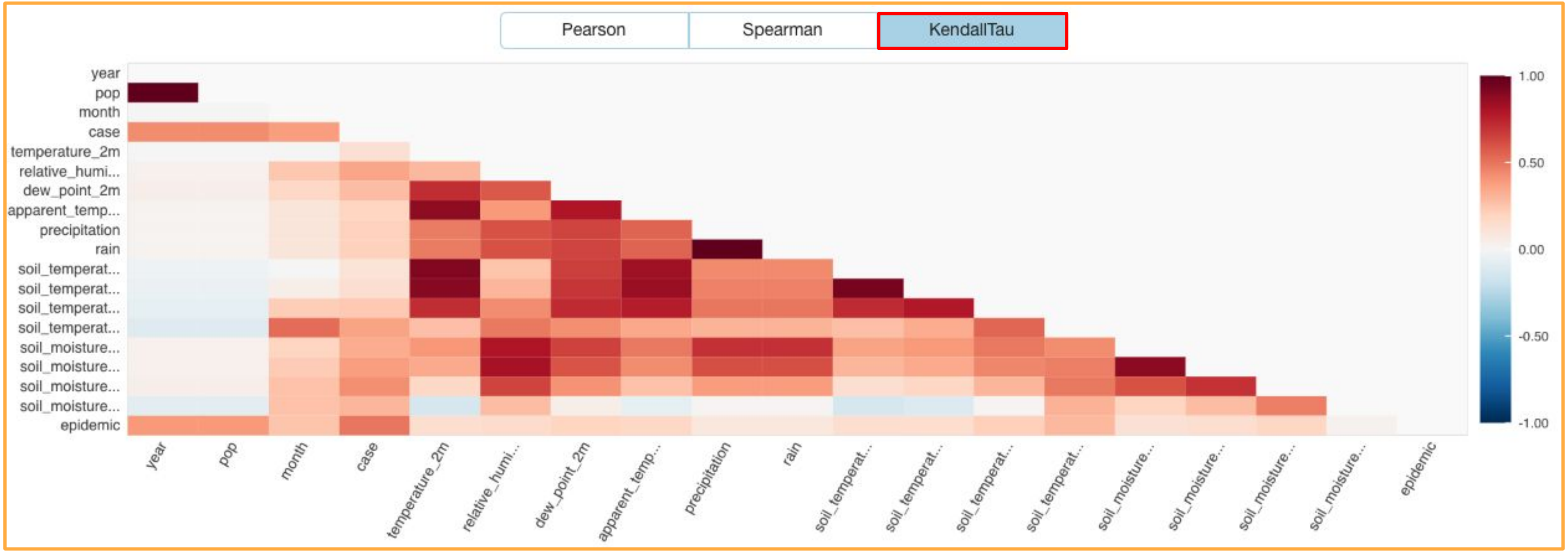
Exploratory Data Analysis

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry



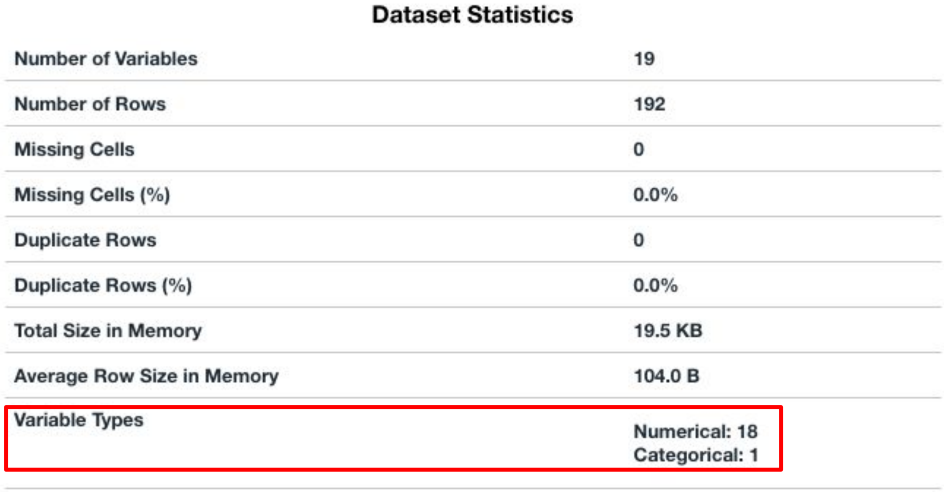
Exploratory Data Analysis

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry



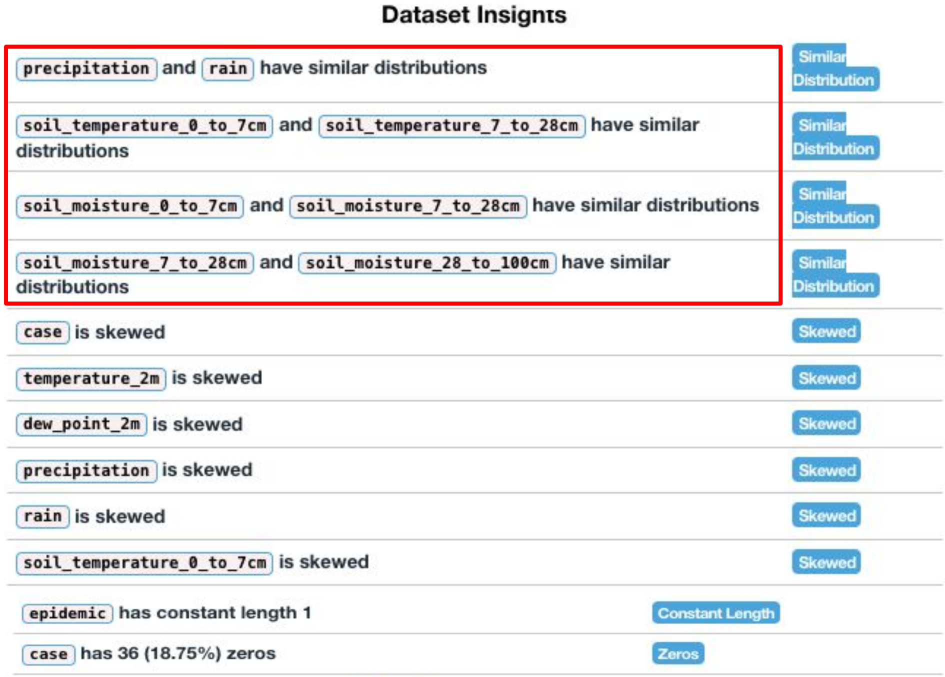
Exploratory Data Analysis

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry



Dataset Statistics

Number of Variables	19
Number of Rows	192
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	19.5 KB
Average Row Size in Memory	104.0 B
Variable Types	Numerical: 18 Categorical: 1

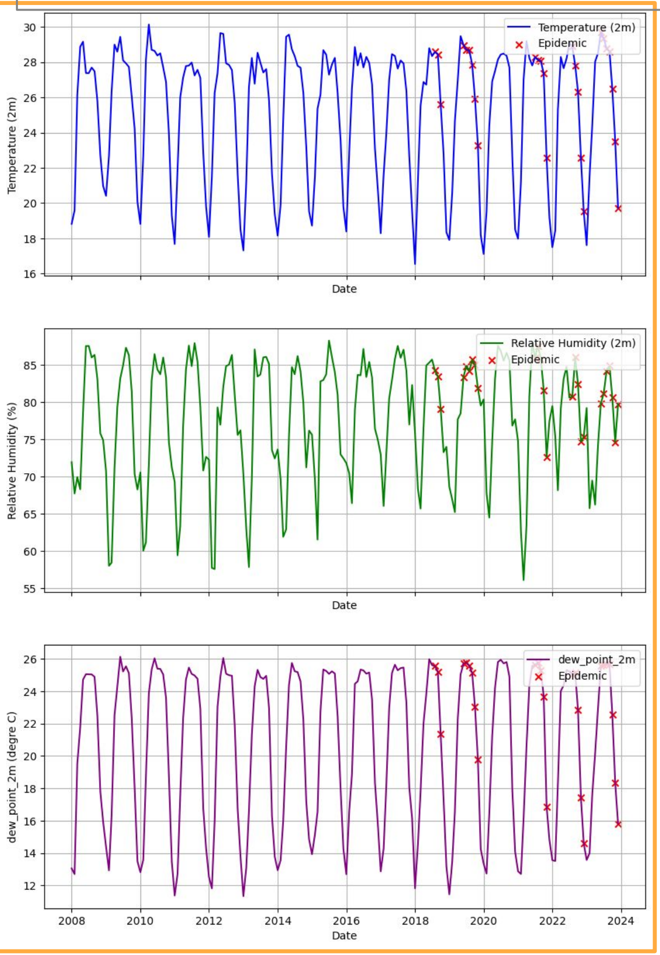


Dataset Insights

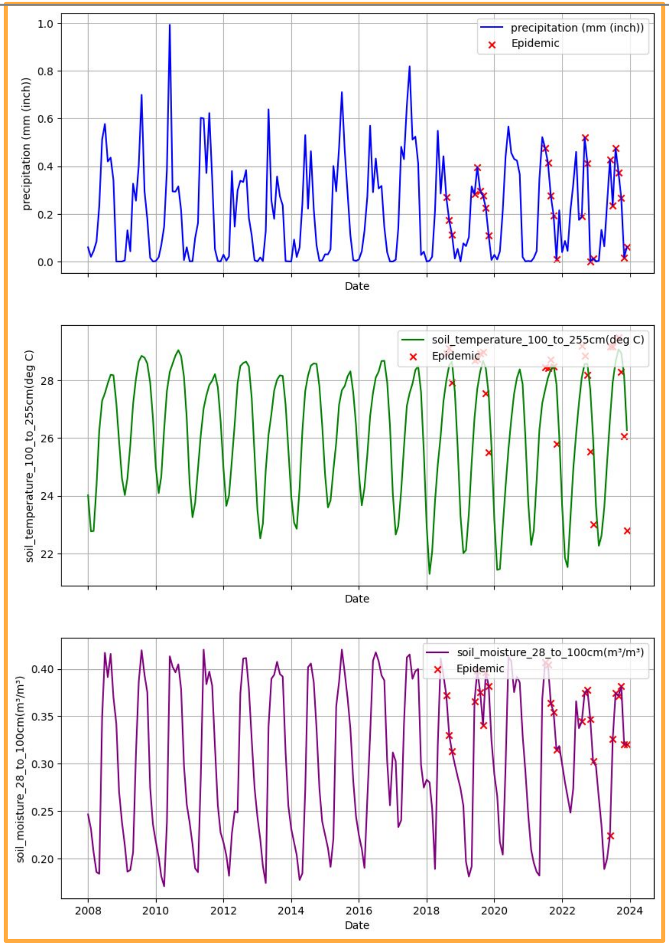
`precipitation` and `rain` have similar distributions	Similar Distribution
`soil_temperature_0_to_7cm` and `soil_temperature_7_to_28cm` have similar distributions	Similar Distribution
`soil_moisture_0_to_7cm` and `soil_moisture_7_to_28cm` have similar distributions	Similar Distribution
`soil_moisture_7_to_28cm` and `soil_moisture_28_to_100cm` have similar distributions	Similar Distribution
`case` is skewed	Skewed
`temperature_2m` is skewed	Skewed
`dew_point_2m` is skewed	Skewed
`precipitation` is skewed	Skewed
`rain` is skewed	Skewed
`soil_temperature_0_to_7cm` is skewed	Skewed
`epidemic` has constant length 1	Constant Length
`case` has 36 (18.75%) zeros	Zeros

Environmental Variables Over Time with Epidemic Indications

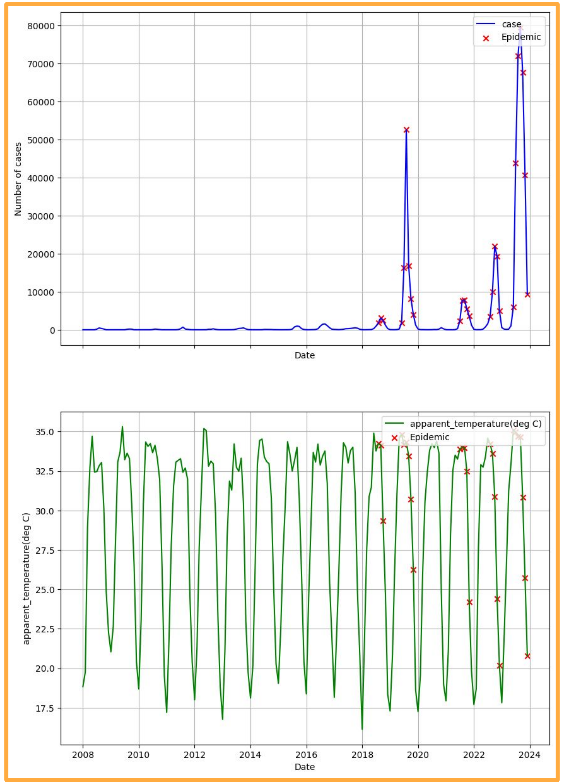
Environmental Variables Over Time with Epidemic Indications



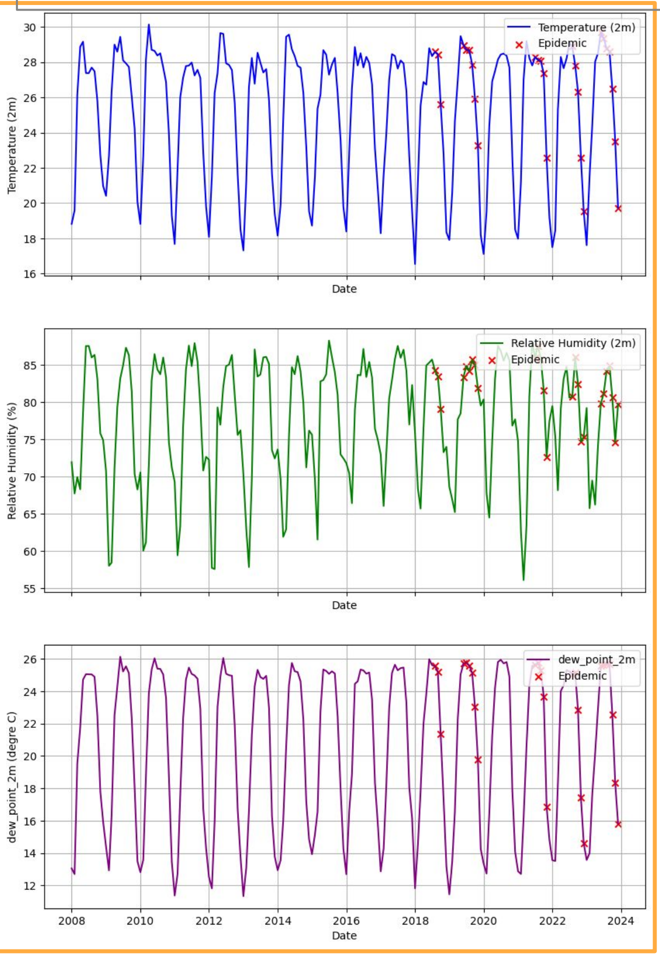
Environmental Variables Over Time with Epidemic Indications

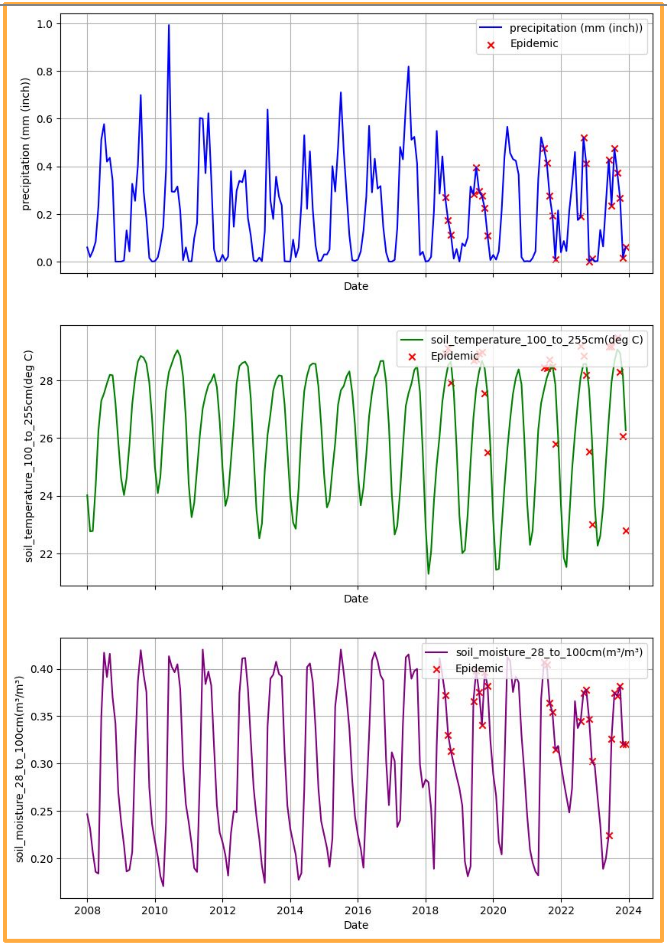


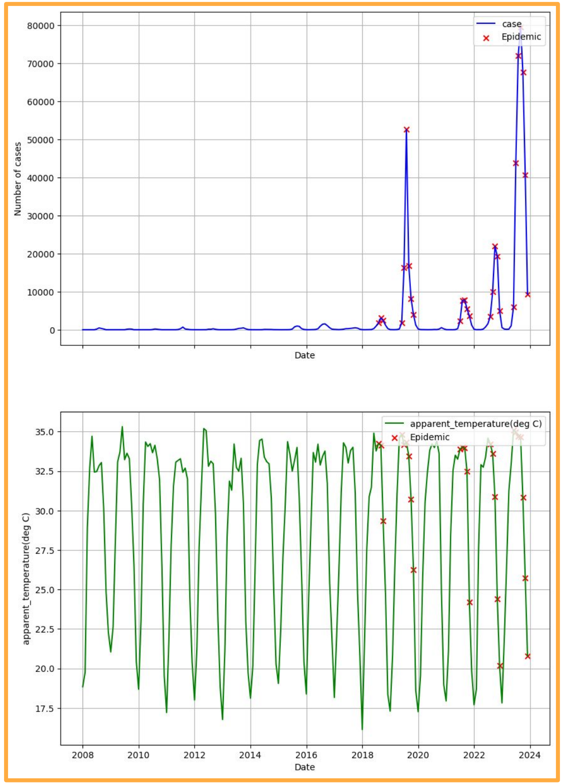
Environmental Variables Over Time with Epidemic Indications



Environmental Variables Over Time with Epidemic Indications



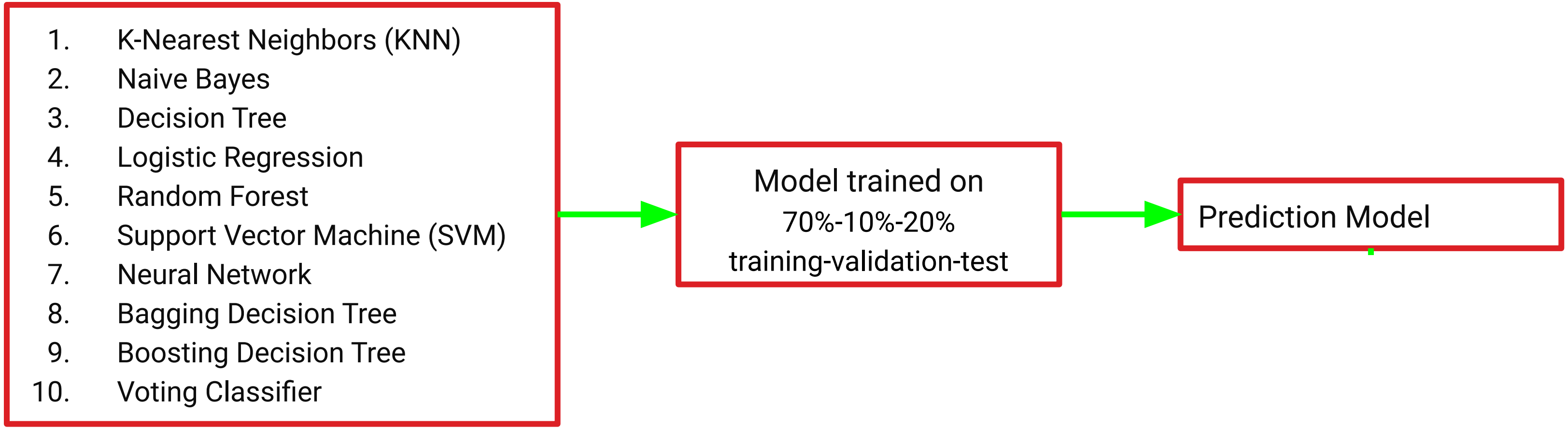




ML Model building

1. K-Nearest Neighbors (KNN)
2. Naive Bayes
3. Decision Tree
4. Logistic Regression
5. Random Forest
6. Support Vector Machine (SVM)
7. Neural Network
8. Bagging Decision Tree
9. Boosting Decision Tree
10. Voting Classifier

ML Model building

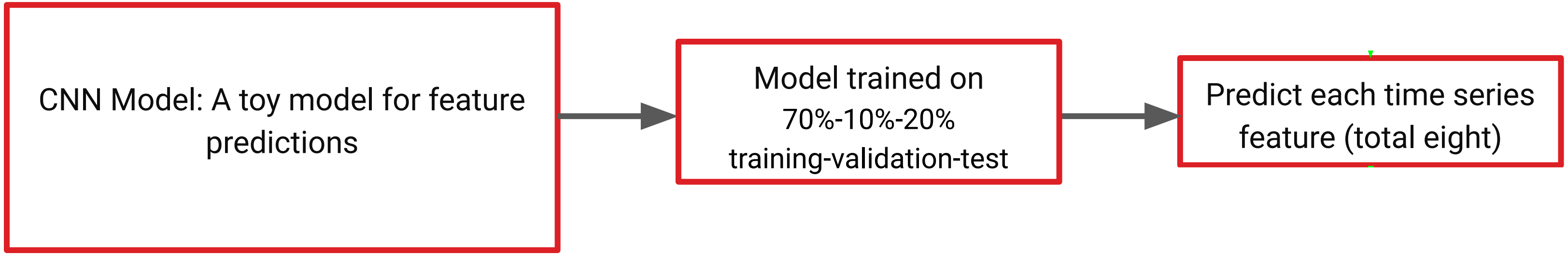


1. K-Nearest Neighbors (KNN)
2. Naive Bayes
3. Decision Tree
4. Logistic Regression
5. Random Forest
6. Support Vector Machine (SVM)
7. Neural Network
8. Bagging Decision Tree
9. Boosting Decision Tree
10. Voting Classifier

Model trained on
70%-10%-20%
training-validation-test

Prediction Model

ML Model building

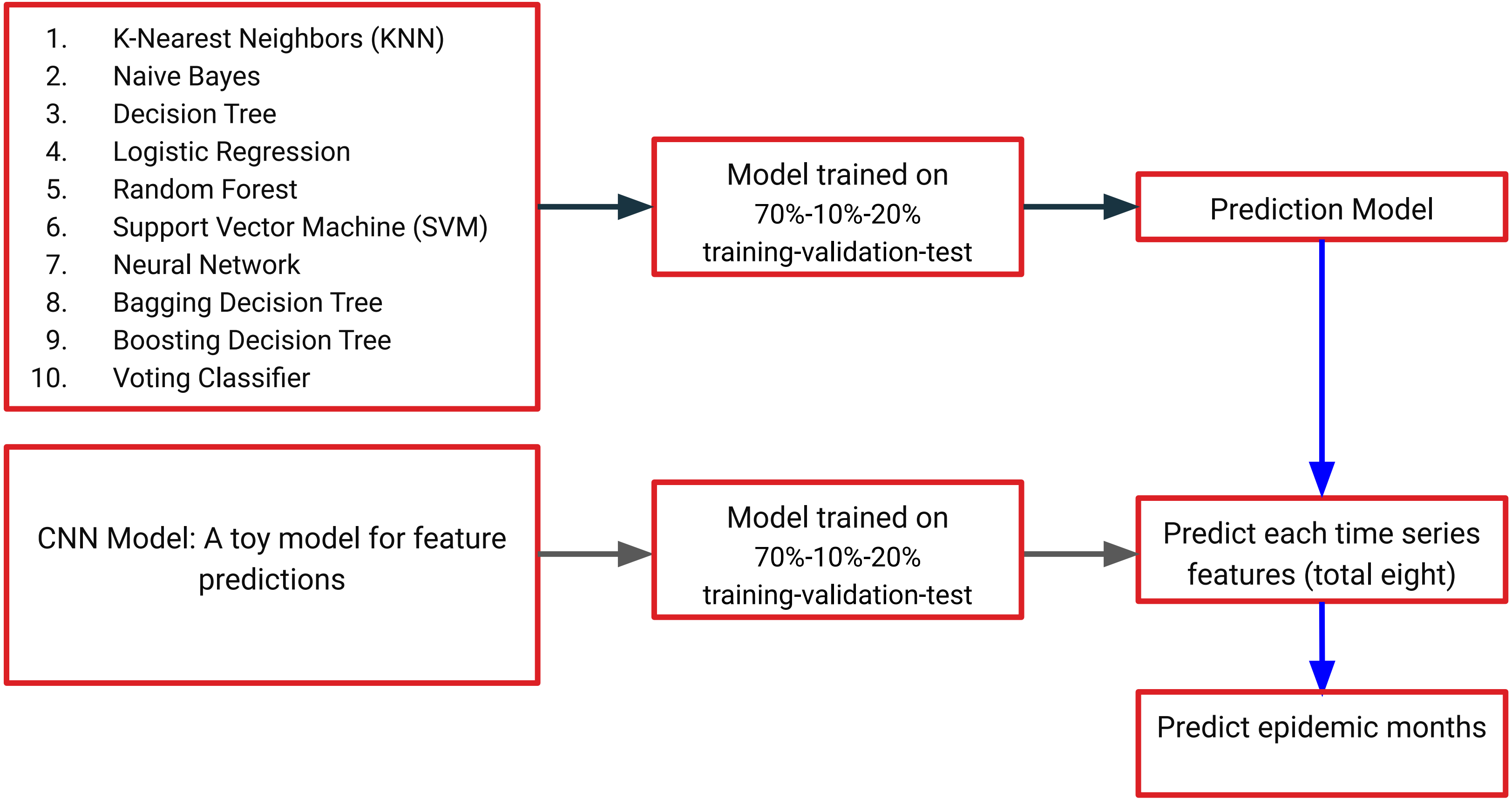


CNN Model: A toy model for feature predictions

Model trained on 70%-10%-20% training-validation-test

Predict each time series feature (total eight)

ML Model building



1. K-Nearest Neighbors (KNN)
2. Naive Bayes
3. Decision Tree
4. Logistic Regression
5. Random Forest
6. Support Vector Machine (SVM)
7. Neural Network
8. Bagging Decision Tree
9. Boosting Decision Tree
10. Voting Classifier

Model trained on 70%-10%-20% training-validation-test

Prediction Model

CNN Model: A toy model for feature predictions

Model trained on 70%-10%-20% training-validation-test

Predict each time series features (total eight)

Predict epidemic months

Model performance

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

THE ERDŐS INSTITUTE
Revolutionary Collaborations in Academia and Industry

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry

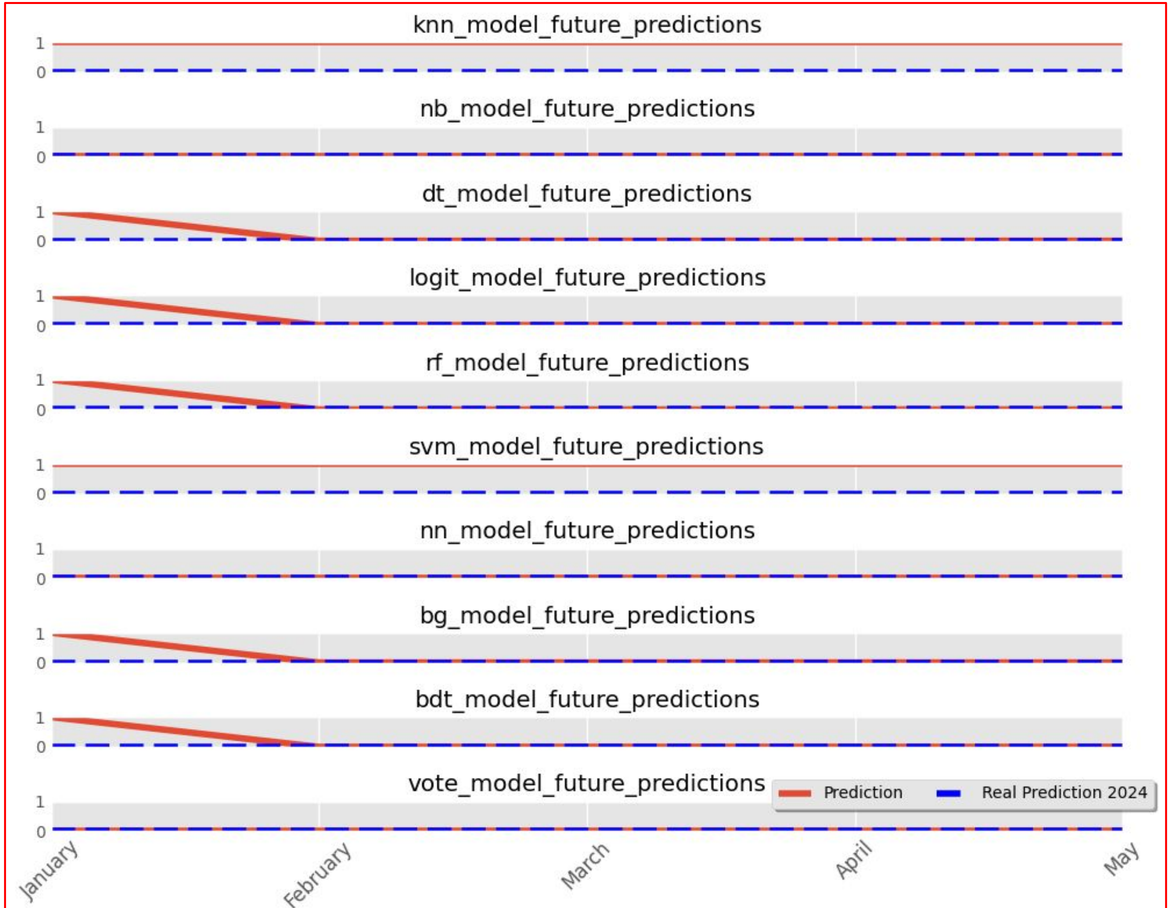
ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model performance

ML model name	Test accuracy	Pred. accuracy	Comments
K-Nearest Neighbors (KNN)	0.95	0.0	Overfitting
Naive Bayes	0.87	1.0	Best model performance
Decision Tree	0.87	0.80	Performed okay
Logistic Regression	0.87	0.80	Performed okay
Random Forest	0.87	0.80	Performed okay
Support Vector Machine (SVM)	1.0	0.0	Overfitting
Neural Network	1.0	1.0	Best model performance
Bagging Decision Tree	0.87	0.80	Performed okay
Boosting Decision Tree	0.87	0.80	Performed okay
Voting Classifier	0.87	1.0	Best model performance

Model prediction

THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry



Modeling limitation

- One limitation in modeling is the potential for the model to overlook complex relationships in the climate data, leading to less accurate predictions in the machine learning models.

Conclusion and future directions

- Classification can be improved by using a bigger dataset.
- Future work will involve implementing sophisticated probabilistic time series forecasting algorithms.

Thank you for joining!

haridas.das@okstate.edu
Department of Mathematics
Oklahoma State University

ahelal@okstate.edu
Department of Mathematics
Oklahoma State University