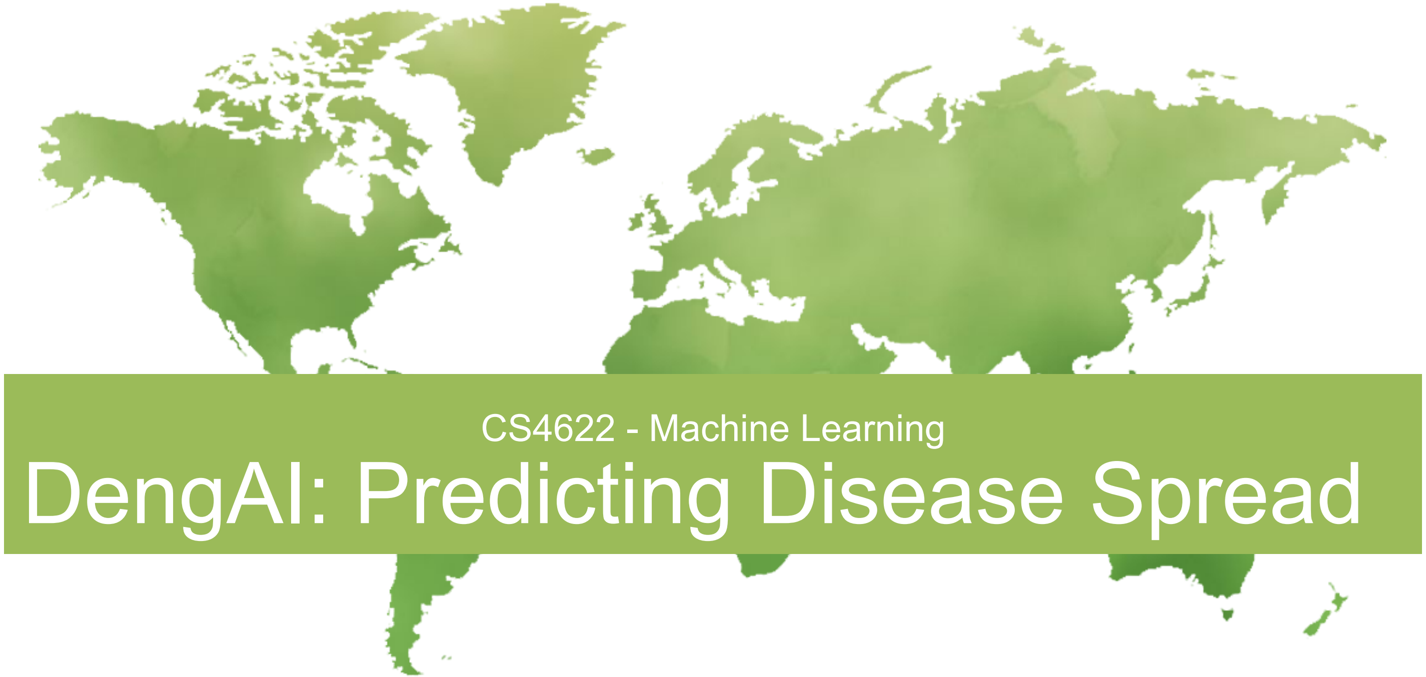


CS4622 - Machine Learning

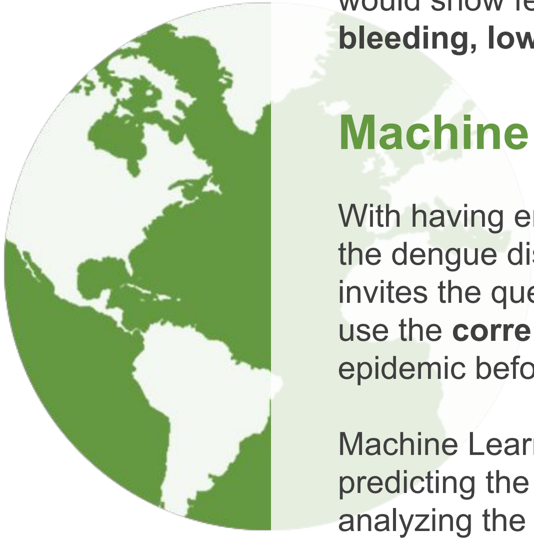
DengAI: Predicting Disease Spread

Team Members:

140004E	Abeygunawardhana H. A. W.
140263U	Jayatilake S. A. P.
140462E	Perera P. D. I. T. S. K.
140623B	Thewa Hettige S. P.

INTRODUCTION

Dengue fever is a deadly viral disease transmitted by mosquitoes. Upon infection, the patients would show fever, rash, and muscle and joint pain. **In severe cases it could cause severe bleeding, low blood pressure, and even death**.



Machine Learning for Dengue

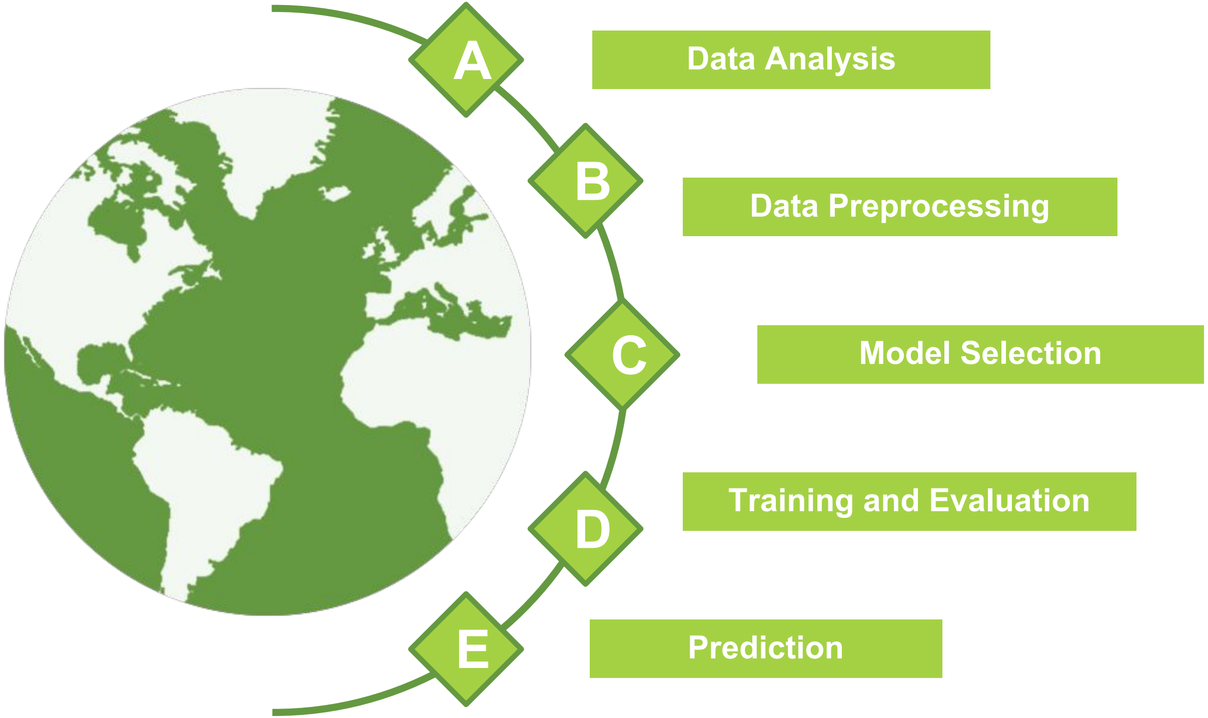
With having enough **historical data** related to the dengue disease and the climate changes, it invites the question of whether it is **possible** to use the **correlation of the data** to predict an epidemic before it happens.

Machine Learning is an efficient way of predicting the future disease outbreaks by analyzing the past data and correlating them with climate changes.

DengAI: Competition

A competition hosted by DrivenData Foundation. The competition is about “predicting the number of dengue cases each week.

Our Methodology



A Data Analysis

B Data Preprocessing

C Model Selection

D Training and Evaluation

E Prediction

The Dataset

The dataset consists of climate data and the total number of reported dengue patients in a weekly timescale

01

City and date indicators

This data describes the city that's related to the data, and the starting date of the week of the data

02

NOAA's GHCN daily climate data

An integrated database of daily climate reports from land surface stations across the globe

03

PERSIANN precipitation readings

PERSIANN-CDR is a daily rainfall estimate at a spatial resolution of 0.25 degrees in the latitude

04

NOAA's NCEP Reanalysis data

Data about how weather and climate are changing over the time

05

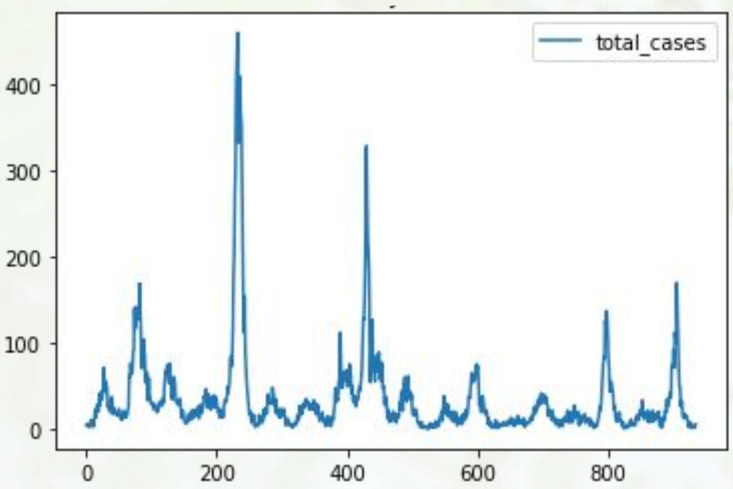
Normalized Difference Vegetation Index

vegetation in a given area, by measuring the difference between near-infrared which vegetation strongly reflects

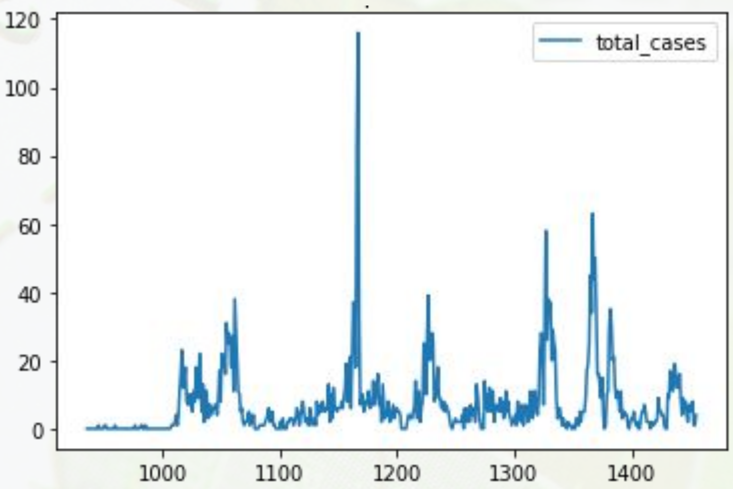
Data Analysis

Raw Data

San Juan



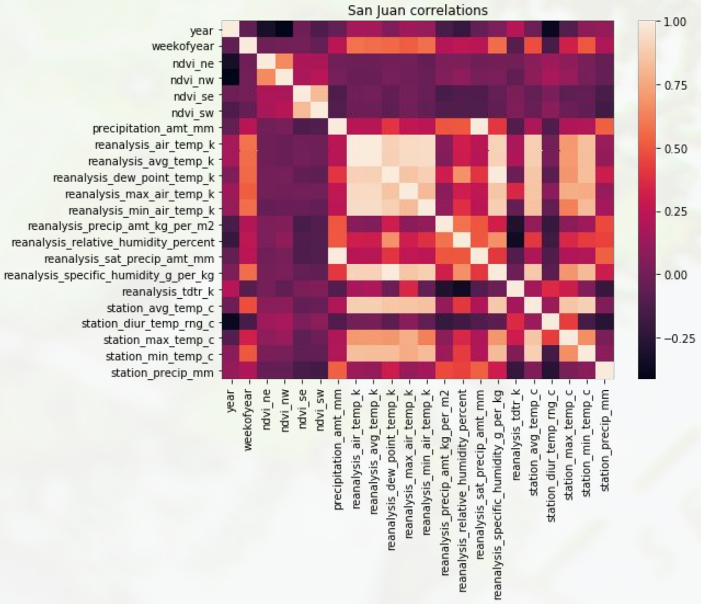
Iquitos



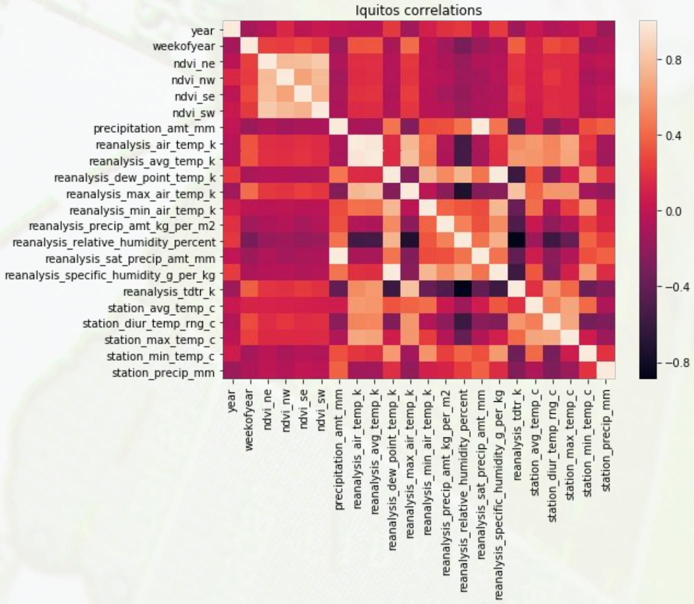
Correlations Among Data

Raw Data

San Juan



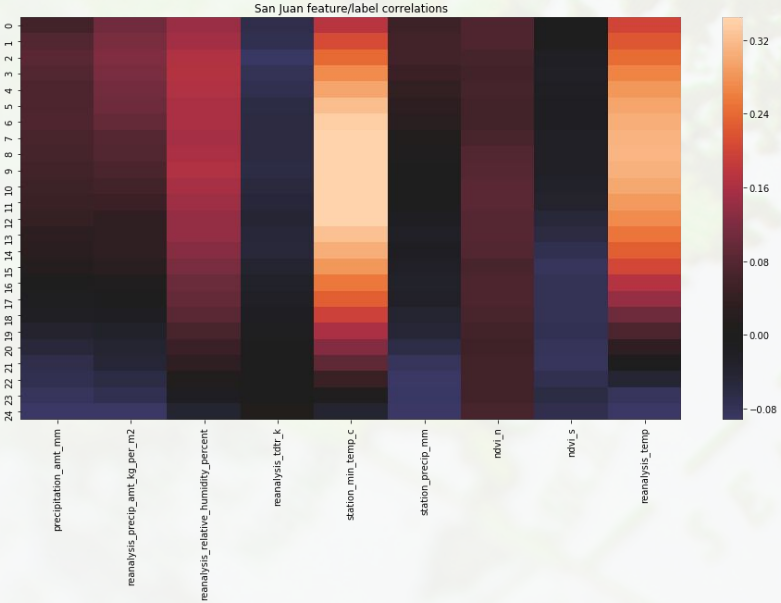
Iquitos



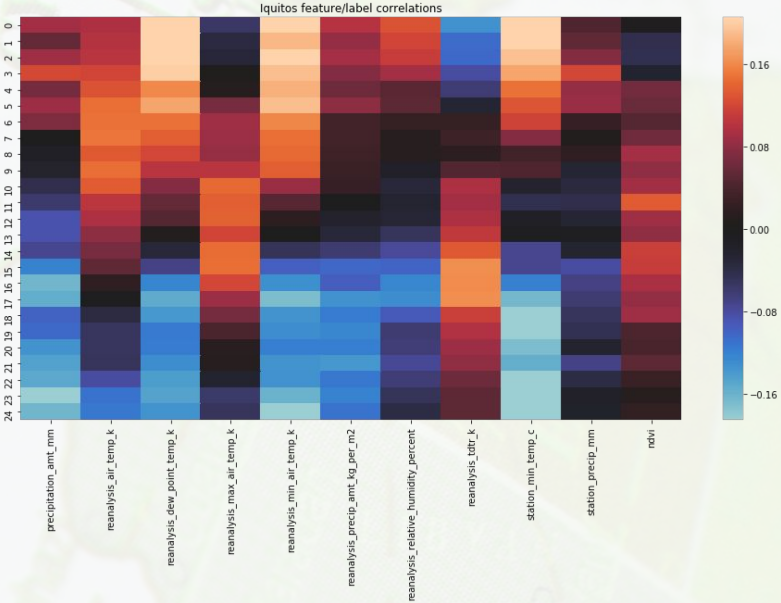
Correlations Among Data

With Time Lag

San Juan

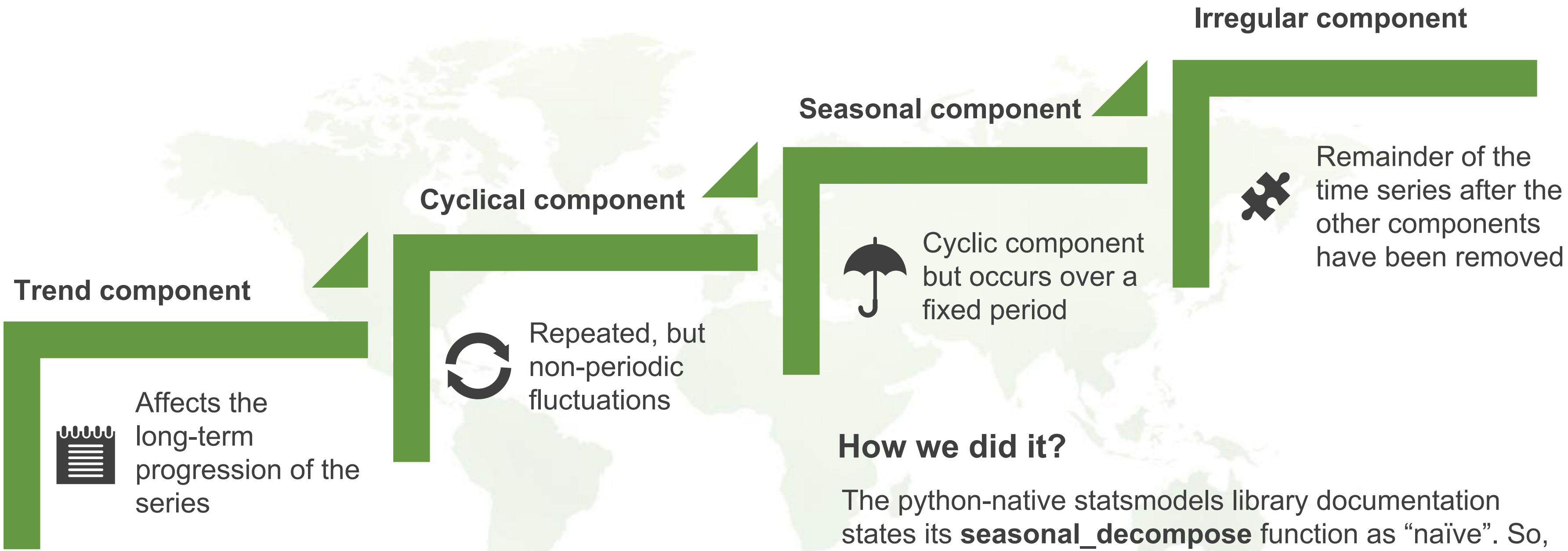

San Juan feature/label correlations

Iquitos


Iquitos feature/label correlations

Time Series Analysis

Statistical time series analysis to separate seasonal, trend and residual components



Trend component

Affects the long-term progression of the series

Cyclical component

Repeated, but non-periodic fluctuations

Seasonal component

Cyclic component but occurs over a fixed period

Irregular component

Remainder of the time series after the other components have been removed

How we did it?

The python-native statsmodels library documentation states its **seasonal_decompose** function as “naïve”. So, we borrowed R’s “robust” STL **(Seasonal and Trend decomposition using Loess)** decomposition function to python using r2py bridge.

Data Preprocessing

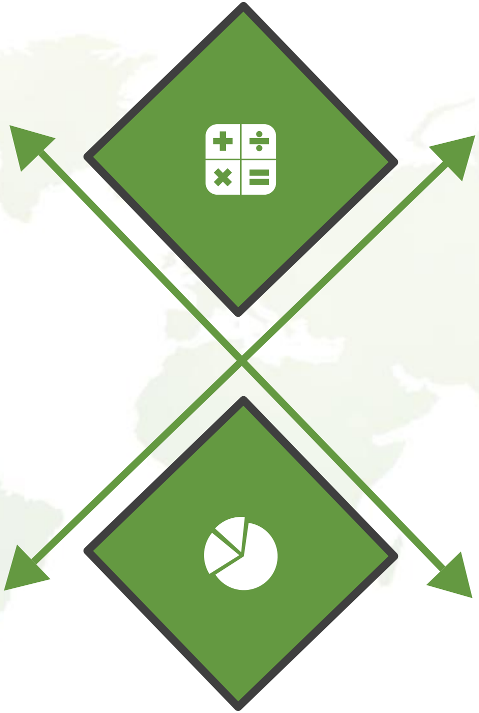
Imputation of missing values

In San Juan dataset, most missing values were NDVIs while in the Iquitos dataset most missing values were station temperatures

Identify outliers and/or smooth

Moving averages were used most of the time to smooth-out outliers. Although median is robust at removing outliers, it removes a significant amount of important “spikes” in data, reducing the impact. Also, the graphs of median-smoothed data look jagged

Data Transformation



Normalization

All Temperature Units were transformed into Celsius scale

Z score of the data was used in the Neural Network based models

Attribute Construction

For some models, we used STL decomposed trend and seasonality components skipping residuals.

Models Used

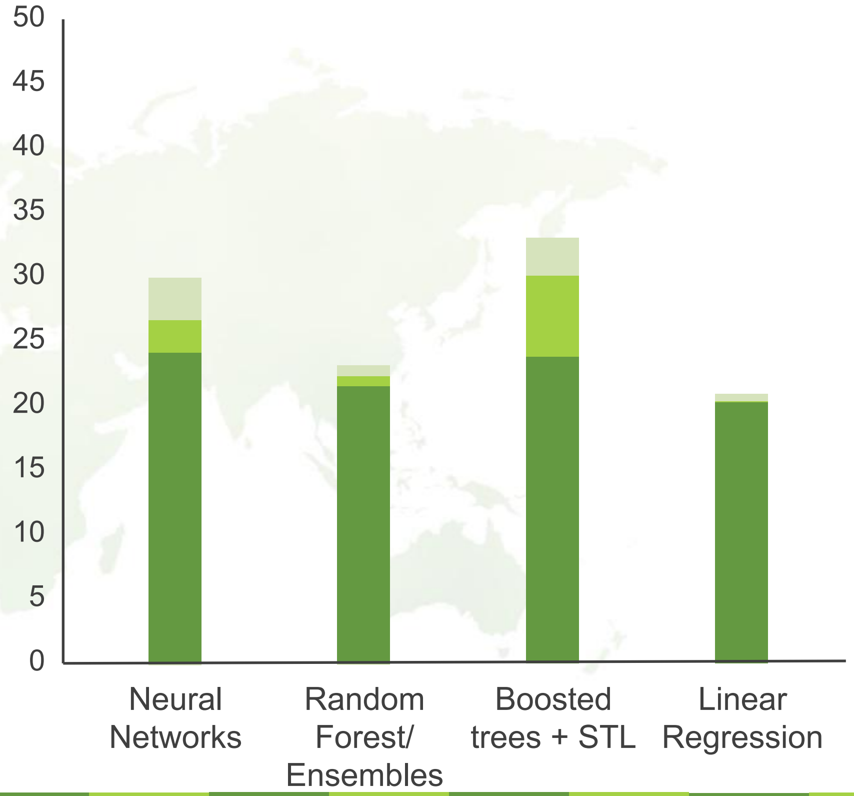
Various models used for evaluation and their approximate results

01 Deep Neural Networks based models

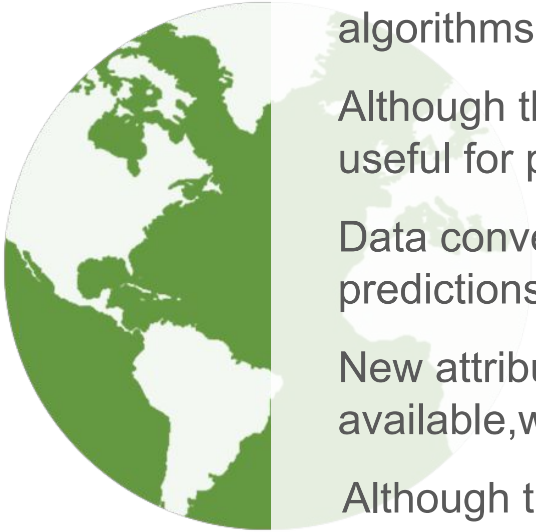
02 Random Forest algorithm and Boosted Trees

03 Boosted Trees with STL

04 Linear Regression



Conclusion



Data preprocessing has a huge impact on the results of machine learning algorithms

Although there are a huge number of attributes, only few of them will be useful for predicting the results

Data conversion and normalization is required for more accurate predictions

New attributes can be created using the attributes that are already available,which will increase the accuracy

Although there are complex and sophisticated models to do regression, simple models are less likely to overfit for the unseen data

Thank You!!!

Any Questions?