Logo

Description automatically generated

PATTERN BASED CHANGE DETECTION ALGORITHMS APPLIED TO CLIMATE DATA​

CRISP-DM

Computer Science

A.Y. 2022/2023

Lorenzo Capocchiano

Domenico Narracci

Contents

[Business Understanding 3](#_Toc122383979)

[Business Objectives 3](#_Toc122383980)

[Business Success Criteria 3](#_Toc122383981)

[Assess Situation 3](#_Toc122383982)

[Data Mining Goal 3](#_Toc122383983)

[Data Mining Success Criteria 3](#_Toc122383984)

[Project Plan 3](#_Toc122383985)

[Data Understanding 4](#_Toc122383986)

[Collect Initial Data 4](#_Toc122383987)

[Data Description Report 4](#_Toc122383988)

[Verify Data Quality 4](#_Toc122383989)

[Explore Data 4](#_Toc122383990)

[Data Preparation 5](#_Toc122383991)

[Select Data 5](#_Toc122383992)

[Clean Data 5](#_Toc122383993)

[Construct Data 5](#_Toc122383994)

[Integrated Data 5](#_Toc122383995)

[Format Data 5](#_Toc122383996)

[Modelling 6](#_Toc122383997)

[Evaluation 6](#_Toc122383998)

# Business Understanding

## Business Objectives

Retrieve and analyze climate data over last decades​ in order to discover patterns and detect how and to what extent climate has been changed over the time​.

The goal of this works is to understand:

1. The period of the year most affected by climate changes
2. Which years are more affected by the change

The reason of this business needs is understanding when and how the climate is changing from starts of 21st century to nowadays.

## Business Success Criteria

Detect Concept Drifts in order to improve predictions on climate data.

## Assess Situation

* **Resources available**:
  + Our PCs with Docker containers
  + Microsoft Teams
* **Data Available**:
  + NOAA
* **Time constraints**:
  + Availability to develop the solution only during nights after workdays and weekends.

## Data Mining Goal

The goal is to mine patterns to be compared and to discover changes inside them.

## Data Mining Success Criteria

The data mining success criteria is identified from the discovery of changes with a minimum support of **0.25**.

## Project Plan

* Selected technologies:
  + Java 8
  + Docker
  + jKarma
  + Spark
  + GitHub

# Data Understanding

The used dataset was completely delivered from **NOAA** (*National Oceanic and Atmospheric Administration*) website, which is one of the most important datacenters for the environmental information.

As said into the business objectives paragraph, data interval is from 2000 to 2022 of climate detections in Louisiana and Mississippi, but they can be changed or extended based on needs.

## Collect Initial Data

As said, the dataset was collected from [NOAA website](https://www.ncei.noaa.gov/cdo-web/).

## Data Description Report

Dataset structure for the case study is composed by:

* **23** CSV Files
* **9** columns
  + **7** numerical
  + **2** categorical (*Station and Name*)

### Verify Data Quality

* **Data Completeness**: some columns have missing data.
* **Data Consistency**:
  + In some rows, numerical columns have ‘9999’ value.

### Explore Data

Descriptions of data:

* **STATION –** Identifier
* **NAME -** (max 50 characters) is the name of the station (usually city/airport name).
* **DATE** - Date of the record in the format *‘yyyy-MM-dd’*
* **AWND** - Average daily wind speed *m/s*
* **PRCP** - Precipitation in *mm*
* **SNOW** - Snowfall in *mm*
* **TMAX** - Maximum temperature in Celsius degrees
* **TMIN** - Minimum temperature in Celsius degrees
* **TOBS** - Temperature at the time of observation in Celsius degrees

# Data Preparation

## Select Data

Drop columns that give no useful information:

* **AWND**
* **TOBS** where ‘*null’*

## Clean Data

* Removed null values
* Sorted values by DATE.

## Construct Data

A new column, named ‘*period’*, was constructed in order to compare rows from the same month in different years.

Period column is a new categorical field which generalize the information of the incoming month for the own row.

## Integrated Data

Phase skipped because there aren’t other data sources available.

## Format Data

* DATE field formatted for visualization scopes removing timestamp information.

# Modelling

The model identified to achieve objectives is based on jKarma, especially in the application of PBCD algorithm with a cumulative sliding strategy.

Those characteristics are performed and completed by the realization of a custom climate data joiner and the applying of an Eclat algorithm with a DFS mining strategy.

In order to improve the speed of all this technical instruments, we parallelized the execution of the workflow using Spark.

# Evaluation