IOWA

Data Analytics

WaterSoftHack ‘24

Objectives

- Perform further tasks for data cleaning and handling missing values, correcting errors.
- Data transformation and extraction from common datasets.
- Perform basic and (advanced) statistical analysis for interpreting data
- Preparing data for standardization, normalization

Tools and Libraries

Web-Based Libraries for Data Analysis

- HydroLang.js, HydroCompute, HydroRTC
- Python for data manipulation and analysis

Tools for Development

- VSCode
- Online Resource (Stackblitz for JS)
- Google Colab (Python)

Links for development

Tutorial links (JS):

HydroLang: https://hydroinformatics.uiowa.edu/tutorials/hydrolang/

HydroRTC: https://hydroinformatics.uiowa.edu/tutorials/hydrortc/

HydroCompute:
https://hydroinformatics.uiowa.edu/tutorials/hydrocompute/

Other links (JS and Python)

Google Earth Engine: https://earthengine.google.com/

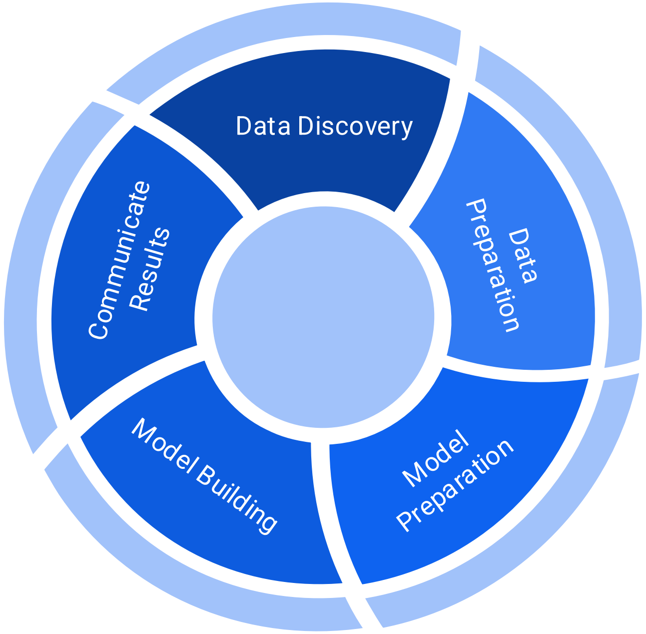
Google Colab (Python) : https://colab.research.google.com/

Part 1 - Introduction

IOWA

Introduction

- Data analytics involve extracting **relevant features** from a set of data
- Data-driven decision-making makes better, robust models that are scalable and customizable



IOWA

Significance

- Ensure data integrity and reliability
- Impact of clean data on analysis accuracy and model performance
- Can simplify the process of data integration for the end use case

Part 2 - Cleaning and Sorting

Common Techniques

- **Handling Missing Values** through imputation, removal, interpolation
- **Removing Duplicates** by identifying and removing any repeated values that shouldn't be there
- **Correcting Errors** particular to the datasets

Imputation

Process of replacing missing data with substituted values

Types

- **Mean/Median**
- **Mode**
- **Regression**
- **K-Nearest Neighbors**

Advantages

- Maintains dataset size
- Simple to implement
- Can improve model accuracy

Disadvantages

- Can introduce bias
- Reduce variability
- Not reflect turthfulness

Removal

Eliminating data entries that contain missing values

Types

- **List Deletion:** removing entire records with missing values
- **Pairwise Deletion:** using only pairs for analysis

Advantages

- Simple, ensuring data integrity for the remaining dataset

Disadvantages

- Can lead to significant data loss
- Reduced sample size
- Bias introduction

Interpolation

Estimating missing values within a range of known data points

Types

- **Linear:** using straight lines between two known datapoints
- **Spline and Polynomial:** fitting within piecewise polynomials
- **Time series Specific:** backward fill, forward fill, seasonal decomposition

Advantages

- Can provide better estimates than simple imputation

Disadvantages

- Requires reliable underlying trends
- Can be computationally intensive

IOWA

Correcting Errors

Identifying inaccuracies or inconsistencies in the dataset

Detecting Errors

- **Validation Rules:** setting constraints and rules for out-of-bounds or incorrect data
- **Data Auditing:** systematic reviews for anomalies
- **Statistical Methods:** tests and algorithms for checking outliers and anomalies, Z-scores and standard deviation checks
- **Visualization:** plotting for pattern identification
- **Cross-Validation:** comparing across sources

Correcting Errors

Fixing Errors

- **Manual Correction:** reviewing and fixing on small datasets
- **Automated Correction:** correction through predefined rules
- **Standardization:** ensure data follows consistent format and structure
- **Normalization:** adjusting values to a common scale
- **Recollection:** regathering data if possible

Part 2 - Preprocessing Techniques

IOWA

Importance

- Ensures **consistent data** for analysis
- Improves **model performance** by aligning data scales and formats
- Handles **missing values, outliers,** and **categorical data**

Techniques

- **Standardization:** scale data to have a mean of 0 and standard deviation of 1
- **Normalization:** rescaling data to a fixed range (e.g. 0 to 1)
- **Encoding:** converting categorical data into a numerical format

Examples:

- Standardizing units of measurement (runoff, evapotranspiration, drought)

Standardization

Scaling data to have a mean of 0 and standard deviation of 1

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

Use Cases

- Features have different units or ranges
- Common in normally distributed data algorithms

Normalization

Rescaling data to a fixed range, typically between 0 to 1.

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Use Cases

- Useful to compute distances between points
- Applied in k-nearest neighbors and neural networks
- Ensures no single feature dominates due to its scale

Encoding

Converting categorical data into numerical format

Techniques

- **One-Hot Encoding:** binary columns for each category

$$\text{Color} = [\text{Red}, \text{Blue}, \text{Green}] \quad \Rightarrow \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- **Label Encoding:** assign each category a unique integer

$$\text{Color} = [\text{Red}, \text{Blue}, \text{Green}] \quad \Rightarrow \quad [0, 1, 2]$$

Part 3 - Data Extraction and Feature Engineering

IOWA

Importance

Selecting the most relevant features to improve model performance, reducing complexity, prevents overfitting

Methods

- **Feature Selection -** correlation, mutual information
- **Feature Engineering -** creating new features from existing data

Examples

- Extract features from landuse data (vegetation cover, imperviousness)

Feature Selection

Correlation

- Measures linearity between two variables

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Values ranges from -1 to 1, with absolute high value showing a strong linear relationship

Mutual Information

- Measures information from one variable through another

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

- Values range from 0 to infinity with higher values indicating strong dependency

Feature Engineering

Enhances the predictive power of models through informative inputs

Techniques

- **Transformation:** apply formulas to transform features
- **Interaction Terms:** features that capture interactions between original features
- **Polynomial Features:** generating polynomial terms to capture non-linear relationships
- **Binning:** transform continuous features into categorical
- **Aggregation:** summary features by aggregating information

IOWA

Part 4 - Statistics and Trend Analysis

IOWA

Descriptive Statistics

Summarizes and describes the main features in a dataset

Key Metrics

- **Mean**
- **Median**
- **Standard Deviation**
- **Kurtosis and Skewness**

IOWA

Inferential Statistics

Makes inferences about a population based on a sample

Key Concepts

- **Hypothesis Testing -** null hypothesis (H0), alternative hypothesis (H1), p-value
- **Confidence Intervals -** range of values within which a population parameter is estimated to lie

$$\bar{X} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

IOWA

Timeseries Analysis

Analyzes data points collected or recorded at specific time intervals

Techniques

- **Decomposition:** Splitting data into chunks of trends, seasonal, and residual components

$$Y_t = T_t + S_t + R_t$$

- **AutoRegressive Integrated Moving Averages (ARIMA):** identifying patterns and forecasting

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

IOWA

Part 5 - Model Generation in Hydrology

Overview of Model Generation

Predictive Models

- Focus on forecasting future events
- Often statistical in nature

Simulation Models

- Aim to replicate physical processes
- Understand and predict future behavior under certain constraints

Importance of Models

Risk Assessment

- Flood prediction
- Drought management
- Resource allocation

Decision Support

- Aids stakeholders to take informed decisions

Understanding System Dynamics

- Comprehending interactions in the hydrological cycle

IOWA

Modelling Techniques

Statistical Approaches

- **Regression Models:** linear regression, multiple linear regression, logistic regression
- **Decision Trees:** CART (Classification and Regression Trees)
- **Random Forests:** Ensemble learning

Example: Generating a predictive model for flood risk assessment

Modelling Techniques

Physically-Based Approaches

- **Distributed Models:** SWAT, MIKESHE, HLM
- **Lumped Models:** single unit watershed (SCS-CN method)
- **Hybrid Models:** combined statistical and physically-based approaches

Example: Using HLM for flood prediction

Part 6 - Example Case Studies

Example 1: Flood Prediction

Needed:

- **Data Sources:** rainfall, soil moisture, river discharge
- **Data Pre and post processing:** cleaning, preprocessing, feature engineering, model generation

Data Sources

Rainfall

- **NOAA:** NLDAS, MRMS
- **USGS:** rain gauges
- **GPM:** satellite-based data

Soil Moisture

- **NASA SMAP:** global data
- **USDA NRCS:** soil climate network

Example 1: Flood Prediction

Data Manipulation

- Handle missing values, normalizing data
- Create new features from moisture index, rainfall intensity, and lagged variables

Model Generation

- Fit data into a physically-based model (NWM or commonly used)
- Create a flood model using statistically based approach

Example 2: Water Quality Analysis

Needed:

- **Data Sources:** water quality measurements and weather data
- **Data Pre and post processing:** cleaning, preprocessing, feature engineering, model generation

Data Sources

Rainfall

- **NOAA:** NLDAS, MRMS
- **USGS:** water quality monitoring
- **EPA WQP**

IOWA

Example 2: Water Quality Analysis

Data Manipulation

- Handle missing values, normalizing data
- Identify patterns, relationships, and trends through descriptive statistics, correlation and regression analyses

Model Generation

- Derive insights from long-term change in water quality parameters
- Use time-series analysis, moving averages, seasonal decomposition

Q/A Discussions

Next Hour - Training