# Preprocessing

**Preprocessing Documentation - Improve the Road Safety in Breda**

Ron L. Tabuchov
Mohamed K. A. M. Elshami
Daria Vladutu
Peter Husen

Data Science and Artificial Intelligence, Breda University of Applied Science

# Index

# 1 Background

This document provides overview of the data preprocessing phase of the data science project. It includes details related to the steps taken to in the data preprocessing process and overall project outcomes.

# 2 Goal

Document the data preprocessing efforts to ensure all steps are recorded and transparent. The goal is to provide clear insights into the preprocessing activities.

# 3 Method

Outline and document all preprocessing steps taken, including data cleaning, transformation, and feature engineering. Each step should be detailed to ensure reproducibility and clarity.

# 4 Preprocessing

**ANWB - Data Set:**

- Data Extraction: Queried the data and converted it into a dataframe.
- Datetime Conversion: Transformed the ‘event_start’ column to datetime format, extracted the date, and set it as the index.
- Feature Creation: Created a new feature ‘avg_speed’ by averaging ‘speed_kmh‘ and ‘end_speed_kmh’.
- Data Type Adjustment: Converted the ‘maxwaarde’ column from float64 to int64 for simplicity.

Column Removal:

- Event_end: Removed as event duration is sufficient and the date from ‘event_start’ is used.
- Municipality name: Dropped as the municipality is always Breda.
- Latitude and Longitude: Removed since only the general location is required.
- Speed_kmh and End_speed_kmh: Replaced by the new ‘avg_speed’ feature.
- is_valid: Removed rows where ‘is_valid’ is False and then dropped this column.
- Road_manager_name: Not needed for the project goal.
- Maxwaarde: Removed as average speed is used instead of exact measurements.

Data Sorting and Classification

- Sorting: Sorted the dataframe by year in descending order.
- Incident Classification: Created a new feature to classify incidents based on the ‘incident_severity’ column into categories: low risk, medium risk, and high risk.

Outlier Management:

- Removed incident severity types with low sample sizes.
- Removed outliers in the duration_seconds column.

Normalization:

- Applied log transformation to normalize the ‘duration_seconds’ column.
- Used RobustScaler to normalize the ‘avg_speed’ column.

Visualization:

- Plotted average speed of roads with the highest number of outliers.
- Plotted average duration of outlier incidents by road names.

Key Findings

- Common Incident Types: Speeding incidents were the most common.
- High Incident Road: ‘Graaf Engelbertlaan’ had the highest number of incidents (4561 total), categorized as:
- Accelerating: 25
- Braking: 105
- Harsh Cornering: 1049
- Speeding: 3382
- Peak Incident Month: May 2018 had the highest number of incidents.

**Open Meteo - Data Set:**

This weather dataset contains meteorological data from January 1, 2015, to May 1, 2024, sourced from the Open-Meteo API. The data includes various weather parameters recorded at a specific location in Breda (latitude 51.5904 and longitude 4.7936). The dataset provides insights into weather patterns and trends over nearly a decade.

Data Fields

time: Timestamp of the recorded data, indicating the specific date and hour of observation.

temperature_2m (°C): Temperature measured at 2 meters above ground level, in degrees Celsius.

rain (mm): Amount of rainfall recorded, in millimeters.

snowfall (cm): Amount of snowfall recorded, in centimeters.

snow_depth (m): Depth of snow on the ground, in meters.

weather_code (wmo code): Weather condition code according to World Meteorological Organization (WMO) standards.

WMO weather codes provide standardized information about observed weather conditions. Some examples include:

0: Clear sky

1: Mainly clear, a few clouds

2: Partly cloudy

3: Overcast

45: Fog

51: Light drizzle

53: Moderate drizzle

55: Heavy drizzle

61: Light rain

63: Moderate rain

65: Heavy rain

71: Light snow

73: Moderate snow

75: Heavy snow

95: Thunderstorm, slight or moderate

97: Thunderstorm with hail

These codes are useful for adding another dimension to risk level analysis.

Risk Level Classification:

- Low Risk: Total intensity less than 0.4 cm (4 mm of rain or snow).
- Medium Risk: Total intensity between 0.4 cm and 1.0 cm (10 mm of rain or snow).
- High Risk: Total intensity greater than 1.0 cm (10 mm of rain or snow).

Data Preprocessing Steps

Data Loading:

- Load the dataset using appropriate file handling methods (the missing CSV file in this case).

Datetime Conversion:

- Convert the time column to datetime format for temporal analysis.

Resampling Data:

- Resample the data to daily frequency to simplify the analysis and reduce noise from hourly fluctuations.

Risk Level Calculation:

- Calculate risk levels based on combined daily intensities of rain and snow.
- Create a function to determine the risk level based on WMO weather codes.
- Apply this function to create a new risk_level column based on weather codes.

Comparison of Risk Levels:

- Compare the two risk level columns (risk_level_rain_snow and risk_level_wmo).
- Highlight and count the differences between the two risk levels.
- Provide summary statistics on the differences.

Data Visualization:

- Plot trends and insights derived from the data, such as monthly averages of temperature, rainfall, snowfall, and snow depth.
- By following these steps, the dataset is cleaned, transformed, and prepared for further analysis, ensuring accuracy and relevance for exploring weather patterns and trends in Breda.

## BRON Data Set:

Initially we had the idea to work with the BRON dataset to see if there was a relation between the incidents and actual accidents that had happened. To see how we could use this the first step was to get the column names. For this we used SQL to print the headers of all the columns. We also wanted to figure out which ones had missing values, and what the value structure looked like. Looking at the columns themselves we figured out that most columns simply had numbers assigned to them, and we had to figure out what the numbers meant, so we found a folder containing most of the information regarding the columns as well as a document guiding us on how to use it. A lot of the columns contained a lot less information than we had initially hoped to find when reading the names of the columns. Some data was about whether there was construction on a road, but we want to figure out why a road is a problem perpetually and not temporarily. Other data was about lighting, but the roads of Breda have street lighting so this would not be a problem even at night. Therefore, we had decided to drop using the BRON dataset and focus on the safe driving dataset instead to find out what are roads where a lot of incidents are recorded.

- Retrieved column headers using SQL.
- Identified columns with missing values.
- Explored the structure of values in columns using external documentation.
- Evaluated the relevance and utility of data based on column content.
- Decided to discontinue using the BRON dataset in favour of a dataset more aligned with research needs.

## KNMI

**Remove Unnecessary Columns:**

- Dropped the columns: NAME, latitude, longitude, dr_pws_10, ww_cor_10, ri_pws_10.

**Date and Time Handling:**

- Split the dtg column into separate date and time columns for better temporal analysis.

**Filter Date Range:**

- Filtered the data to include only records between January 1, 2018, and December 31, 2023.

**Handle Missing Values:**

- Removed rows with missing values to ensure data completeness.
- Imputed missing values using appropriate methods:
- Mean
- Median
- Mode
- Fixed value (e.g., 0)

**Convert Object Columns and Interpolation:**

- Converted object type columns to their appropriate data types.
- Interpolated missing values to ensure continuity in the data.

**Outlier Detection:**

Identified outliers in the dr_regenm_10 and ri_regenm_10 columns using the Z-scores method.

**Data Transformation:**

- Applied log transformation to normalize the data distribution.
- Used square root transformation as an alternative normalization method.

**Split Dataset:**

- Divided the dataset into training, validation, and test sets for model development and evaluation.

**Risk Level Feature Creation:**

- Created a new feature to determine the risk level of rain (Low, Mid, High) using the rain intensity and duration columns.

**Visualization:**

- Developed visualizations to explore the data:
- Distribution of risk levels.
- Risk levels over time.