# Assignment 3 - Optimization of Long-Term Correction of Wind Data Using Regression Models

1. Assignment description / Problem statement

Optimize Vesta’s methods for prediction of long-term corrected (LTC) wind data, which are used for planning the locations of new wind farms. Build and fine-tune a linear regression model with reasonable accuracy that can possibly replace accurate but resource-demanding neural network model that is currently serving this purpose.

1. Analysis & Preprocessing

Firstly, two provided datasets were loaded into Pandas data frames variables. The two datasets, meso and mast data, both store wind speed and direction values however in different heights and different time frequencies. While meso data contains hourly measurements in heights of 60, 80, 100, 120 and 140 meters, mast data contains measurements taken every 10 minutes in heights of 44, 77 and 125 meters.

We checked for null values per column and if the non-null values are valid (bigger than 0 for the wind speed columns and between 0 and 360 for the wind direction columns).

To combine the two provided data sources, we had to resample the more frequent mast data and turn them into hourly measurements. To achieve this, we grouped the data within hour frames and aggregated its values, using mean for wind speeds and median for wind directions. Having two data frames with the same granularity we could finally merge them into one. As next we decided to drop rows containing null values. We could either replace nulls with one of the aggregated values (mean or median), take a value from the closest measurement point and replace the null or use some more advanced feature engineering technique (e.g. linear regression). We decided to remove them as there were not as many missing values after the mast data resampling where we eliminated them. After that we checked if the wind speed data fits the Weibull Distribution as we were advised, which they did.

We checked for correlations between wind speed in the height of 77 meters to find out which of the values influence its value the most. Eventually we scaled the data using the StandardScaler and split them into train and test sets.

1. Used algorithms

We found the best alpha and optimal Lasso vs Ridge ratio using the ElasticNet cross validation and then trained the ElasticNet model using the alpha and the ratio achieved in the previous step.

1. Performance

A black background with white numbers

Description automatically generated

1. Reflection / Learning outcomes

Surprisingly enough we were able to achieve significant results. With deep data analysis and proper preprocessing, we ended up with a nice and clean data set that could be used for the final step, model training and tuning.

Due to lack of knowledge about wind data and deeper understanding of this topic we used wrong data features combination for the training of our model which resulted in too good but useless results, as we used measurements from other heights for determining the values in the 77 meters height.