**Assignment 1**

***Description of the forecasting problem***

Weather data for the city of Szeged in Hungary, from 2006 to 2016, is available. This forecasting problem deals with predicting the **Apparent Temperature** value given a value of **Humidity** on that day.

Y = Apparent Temperature

X = Humidity

Where X is the independent variable/attribute, also known as predictor variable. Based on the values of X, we will predict the values of Y which is our variable of interest.

***Description of the available data***

|  |  |  |
| --- | --- | --- |
| **Column name** | **Type** | **Description** |
| Formatted Date | DateTime | Date and Time of the day |
| Summary | String | Short Summary of the day |
| Precip type | String | Type of precipitation |
| Temperature | Numeric | Actual Temperature |
| Apparent Temperature | Numeric | Temperature perceived by humans |
| Humidity | Numeric | Value of humidity |
| Wind Speed | Numeric | Speed of Wind |
| Wind Bearing | Numeric | Direction of the Wind |
| Visibility | Numeric | Distance at which an object or light can be clearly discerned |
| Loud Cover | Numeric | Total cover |
| Pressure | Numeric | Value of atmospheric pressure |
| Daily Summary | String | Overall summary for the day |

**Attributes used:**

Y (Variable of interest) = Apparent Temperature

X = Humidity

***Short overview of the selected algorithms***

**Linear Regression**

Regression – an approach for modeling the relationship between a dependent variable and independent variables

Linear regression – linear relationship between a dependent variable and independent variables

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. If there are more than one predictor variables, the prediction is called *multivariate linear regression.*

**ARIMA (Time series)**

In statistics and econometrics and in particular in time series analysis an **autoregressive integrated moving average (ARIMA)** model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting) ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.