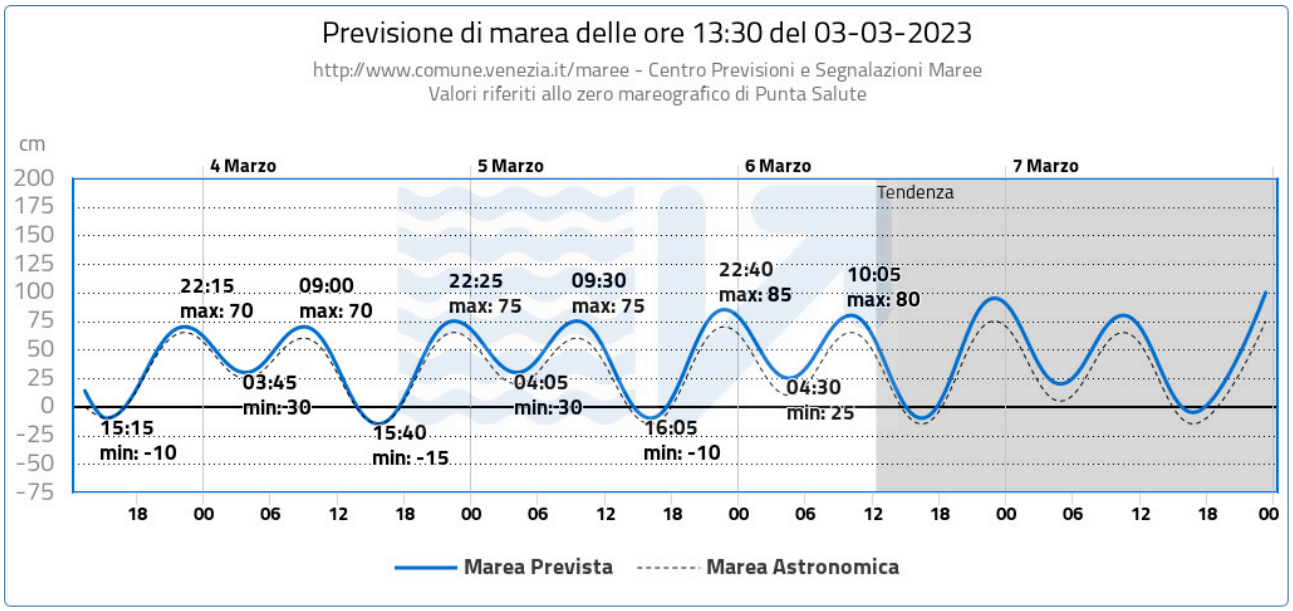
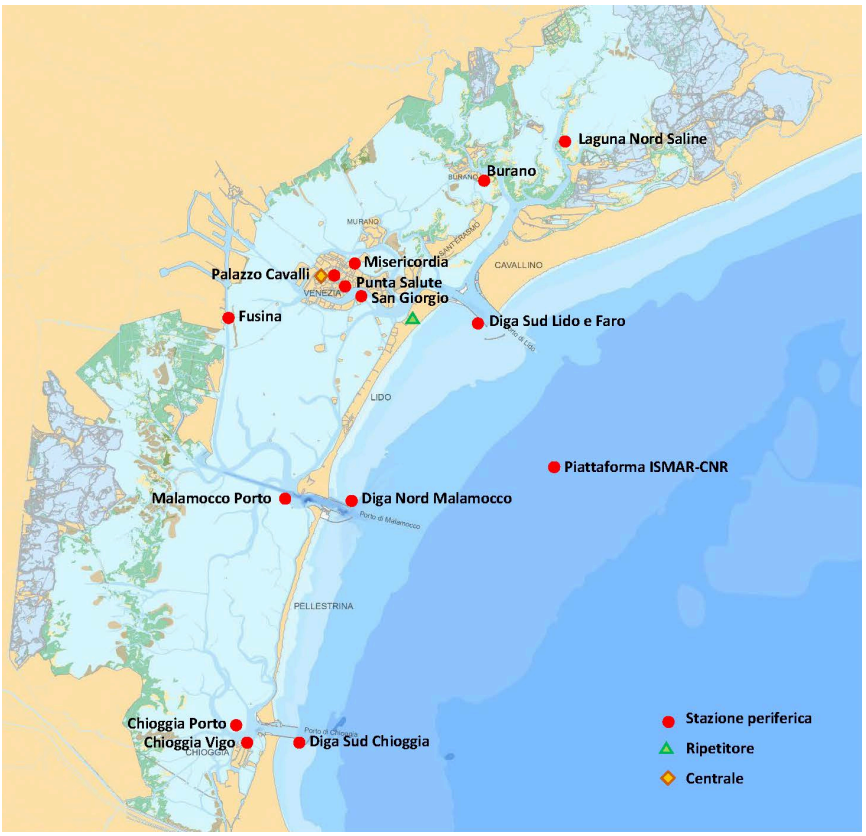
High Water in Venice
Final Data Challenge (Graded)

20596 - MACHINE LEARNING - Prof. Pieralberto Guarniero
Bocconi University - May 5th, 2023

You are provided with a dataset concerning **high tides** in Venice. These data are used by the "*Centro Previsioni e Segnalazioni Maree*" to produce forecasts of high tide.



The analysis of high tides in Venice has a very long history, and this complex procedure is handled by the **"Centro Previsioni e Segnalazioni Maree"**, located in **Palazzo Cavalli**. If you are unfamiliar with this phenomenon, you may want to read this short **booklet (ITA and ENG)** to get an overview.



The data comprises information gathered from a number of Venetian meteorological station. Each station tracks different kinds of information, depending on the sensors that have been installed.

The training dataset consists of 1000 measurements (taken at different times) of input variables that can be used to predict high tide. Note that the observations are not ordered in time. For these observations you know whether the phenomenon of high tide was observed exactly 6 hours following the observation (y = 2) or not (y = 1). For the additional 866 measurements in the test set, you only have the inputs.

Your goal is to predict y for the held–out 866 unit in the test set. There are **37 input variables**, most of which are self explanatory, consisting of measurements from the different stations of the quantities below.

1. Astronomical_Tide+6, the astronomical tide level exactly six hours following the time of the observation (see note on tide below)
2. Average_Tide_Level
3. Average_Wind_Direction, measured in degrees
4. Average_Wind_Speed
5. Max_Wind_Speed
6. Humidity
7. Solar_Radiation
8. Air_Temperature
9. Water_Temperature
10. Pressure
11. Significant_Wave_Height, the average height of the 33% highest waves
12. Max_Wave_Height
13. Rain_Level

Loss function. When miss-classifying by failing to predict an instance of high tide you pay a cost of 5. When miss-classifying by predicting high tide when it does not occur you only pay a cost of 1.

Astronomical tide and high water in Venice. The observed tide in Venice can be thought as the sum of two distinct components: the astronomical tide, correlated to the motion of heavenly bodies, mostly Moon and Sun, and the meteorological surge due to atmospheric conditions. Under normal conditions, the meteorological contribution is small and the observed water level approximately coincides with the one given by the astronomical tide. In case of adverse weather conditions, typically low pressure and strong sirocco winds, meteorological contribution becomes important: if it is on phase with a maximum of astronomical tide, it can produce important phenomenons. The meteorological contribution may also be negative, as a result of a high atmospheric pressure, and thus determine significant low tides. Phenomena such as eustatism (sea level rise) and subsidence (lowering of the soil through natural or anthropogenic causes) gave their strong contribution to the frequency increase of high tide events in Venice during the past decades. Conventionally, in Venice, a sea level higher than +90 cm above the local datum of Punta Salute, is called “acqua alta”: at this level many problems arise involving transport and pedestrian use of roads in lowest areas of the town (Piazza S.Marco). When the tide exceeds +100 cm (5% of public pedestrian land flooded), the phenomenon begins to affect larger sections of pedestrian areas. At an altitude of +110 cm, about 12% of the city is affected by flooding. But when level reaches +140 cm, about 59% of the city is flooded.

Motivation. MOSE (*Modulo Sperimentale Elettromeccanico* i.e. Experimental Electromechanical Module) is an integrated system consisting of rows of mobile gates installed at the Lido, Malamocco, and Chioggia inlets that are able to isolate the Venetian Lagoon temporarily from the Adriatic Sea during acqua alta (high tides). MOSE is designed to protect Venice and the lagoon from tides of up to 3 metres. **The system takes six hours for full activation**. Currently it is raised for tides of more than 110 centimetres.

Time series models can be deployed for the high tide prediction problem. Over a long period the time aspect is significant due to eustatism and subsidence, mentioned above. However over a short period time series, said phenomena are negligible and the time aspects might become irrelevant if the astronomical tide level is known (which will most likely be the most prominent and significant variable for your high water prediction).