ASSIGNMENT 1: MEDICAL COST

Projects form an important part of the education of software engineers. They form an active method of teaching, as defined by Piaget, leading to a "training in self-discipline and voluntary effort", which is important to software engineering professionals. Two purposes served by these projects are: education in professional practice, and outcome-based assessment.

Data cleaning or data scrubbing is one of the most important steps previous to any data decision-making or modelling process. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Data cleaning is the process that removes data that does not belong to the dataset or it is not useful for modelling purposes. Data transformation is the process of converting data from one format or structure into another format. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format. **Essentially, real-world data is messy data and for model building: garbage data in is garbage analysis out**.

This practical assignment belongs to Data Science Master at the UPC, any dataset for modelling purposes should include a first methodological step on **data preparation** about:

- Removing duplicate or irrelevant observations
- Fix structural errors (usually coding errors, trailing blanks in labels, lower/upper case consistency, etc.).
- Check data types. Dates should be coded as such and factors should have level names (if possible, levels have to be set and clarify the variable they belong to). This point is sometimes included under data transformation process. New derived variables are to be produced sometimes scaling and/or normalization (range/shape changes to numeric variables) or category regrouping for factors (nominal/ordinal).
- Filter unwanted outliers. Univariate and multivariate outliers have to be highlighted. Remove register/erase values and set NA for univariate outiers.
- Handle missing data: figure out why the data is missing. Data imputation is to be considered when the aim is modelling (imputation has to be validated).
- Data validation is mixed of 'common sense and sector knowledge': Does the data make sense? Does the data follow the appropriate rules for its field? Does it prove or disprove the working theory, or bring any insight to light? Can you find trends in the data to help you form a new theory? If not, is that because of a data quality issue?

The data set is proposed in Machine Learning with R by Brett Lantzto and can be found in the Kaggle website (https://www.kaggle.com/datasets/mirichoi0218/insurance), there are 1338 individual observations.

Columns	Description
age	age of primary beneficiary
sex	insurance contractor gender, female, male
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
children	Number of children covered by health insurance / Number of dependents
smoker	Smoking
region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
charges	Individual medical costs billed by health Insurance **(target variable)**

The file contains the charges paid as individual medical costs and individual characteristics.

- Create factors for qualitative variables.
- Determine if the response variable (charges) has an acceptably normal distribution.
- Address tests to discard serial correlation.
- Detect univariant and multivariant outliers, errors and missing values (if any) and apply an imputation technique if needed.
- Preliminary exploratory analysis to describe the relationships observed has to be undertaken.
- If you can improve linear relations or limit the effect of influential data, you must consider the suitable transformations for variables.
- Apart from the original factor variables, you can consider other categorical variables that can be defined from categorized numeric variables.
- You must take into account possible interactions between categorical and numerical variables.
- When building the model, you should study the presence of multicollinearity and try to reduce their impact on the model for easier interpretation.
- You should build the model using a technique for selecting variables (removing no significant predictors and/or stepwise selection of the best models).
- The validation of the model has to be done with graphs and / or suitable tests to verify model assumptions.
- You must include the study of unusual and / or influential data.
- The resulting model should be interpreted in terms of the relationships of selected predictors and its effect on the response variable.