Data processing for machine learning

Purpose of the work

Get acquainted with the basic techniques of descriptive data analysis to prepare for machine learning

Tasks to complete
On the website, select a data set for each student individually. It is best to choose datasets in the subject area based on personal preferences. Do not choose too large data sets (preferably no more than 100 MB).
At the beginning of the analysis, a general characteristic of the data set should be given: the number of observations, features, subject area, content of features, the nature of the target variable, the learning task (regression, classification, number of classes).
To carry out a descriptive analysis of the selected set in Jupyter by means of pandas, sklearn, seaborn, including (but not limited to) the following steps:
investigation of the measurement scale of each essential feature;
construction of an empirical distribution of each essential feature;
analysis of the number and distribution of missing values;
construction of a joint distribution of each attribute and target variable;
construction of a correlation matrix.
For each point of the analysis, it is necessary to draw a meaningful conclusion.
Carry out preparatory data processing, including the following steps: 6. removing or filling in missing values; 7. bringing all features to a binary or numeric scale; 8. removing irrelevant or redundant features; 9. other necessary actions depending on the data set and the task (grouping, removing anomalies, etc.).
Quantitative characteristics of the dataset should also be given after its processing.

Security questions
What are the methods for eliminating missing values in the dataset? Based on what it is necessary to apply different methods?
What does the correlation coefficient show? What are its limitations?
What signs can be considered insignificant?
What methods of converting categorical features into numerical ones exist?
What statistical distributions are most often found in real problems?
What is joint and conditional distribution?

Additional tasks
Use multiple machine learning models to solve the task.
Use the method of measuring the training time of each model used.
Generate a summary table of training results, including the following data:
model training time;
accuracy;
precision;
recall;
f1-score.