Home work 5

GitHub repository: https://github.com/Trik17/MachineLearningProject

a) What is your data and task (including data preprocessing steps)?

I started my work from a CSV file taken from Kaggle, it contains information about the answers given from 1010 student (rows of the database) to 150 questions (features, columns). Among these questions there is one in particular “I am empathetic person.: Strongly disagree 1-2-3-4-5 Strongly agree” that is what I want to predict using the answers given to the other questions, I’m only interested in predicting if a student is very empathetic (value of 4-5) or not very empathetic (1-3). The database contains missing values, so the first thing that I’ve done is to deal with them. I’ve eliminated the rows where the feature “empathy” was missing because those data are not useful for neither training or testing. For the other missing values, I’ve imputed them using the mean value of each feature for the integer ones (because most of the algorithms uses them as floats) and the mode for the categorical values. Then, in order to deal with the categorical features that can’t be used from most of the algorithms I’ve transformed them into numerical ones. I used one-hot-encoding for the categorical features “Genger”, “Left\_right\_handed”, “Only\_child”, “Village\_town”, “House\_block\_of\_flats” that represents actual categorical characteristics of a student. For the others, I’ve transformed them in numerical features manually assigning at each value a number, because they actually represent scales of value, like the Smoking attribute where the scale start with “never smoked” and end to “actual smoker”, so using one-hot-encoding is not correct because it will lead to lose this “scale” of values.

(b) what ML solution did you choose and, most importantly, why was this an appropriate choice?

(c) how did you choose to evaluate success, including baselines, experimental setup (e.g., %

train/dev/test), metrics?

(d) what software did you use and why did you choose it?

(e) what are the results?

(f) show some examples from the development data that your approach got correct and some it got

wrong: if you were to try to fix the ones it got wrong, what would you do?