Homework 1
Classification & Regression

University of Bucharest, Faculty of Mathematics and Informatics
Sparktech Software

October 24, 2018

Due date: November 21, 2018

1 Introduction

You'll need to use the algorithms studied so far to make predictions based on the *Dog Breeds Dataset*. This dataset contains the following columns: *Weight(g), Height(cm), Energy level, Attention Needs,Owner Name, Coat Length, Sex, Breed Name, Longevity(yrs)*. Some of these attributes are numerical, some are categorical. You will have to do Classification and Regression and compare different models with different hyper-parameters. ***Breed Name, Longevity(yrs)*** are the **labels** you need to predict. You can use all the other columns as **features**. More details of each task are presented in their appropriate section.

The models you need to explore and compare for this problem are the following:

1. Linear Regression (Ridge, Lasso)
2. Logistic Regression
3. Random Forests
4. KNN

The goal of this assignment is to find the best hyper-parameters for each model, and the best model for each task. You'll need to detail the steps you took towards reaching your conclusion, by writing and submitting a report, either as a *Jupyter Notebook* or a *PDF*. Note that you also have to submit your source code, even if you choose to submit a *PDF*.

2 Task 1: Classification

For this task, you have to find the best model for predicting the *Dog Breed* label. For this, use any features you like (even features you engineer yourself, except *Longevity*).

Make sure to include proper *matplotlib* visualizations (e.g. decision boundaries) and tables to support your analysis. Also, make sure your model is ***properly evaluated*** with the appropriate metrics and evaluation sets.

Your analysis must be a comparison between *Logistic Regression, Random Forests* and *KNN*.

3 Task 2: Regression

For this task, you have to find the best model for predicting the *Longevity* label. For this, use any features you like (even features you engineer yourself, except *Dog Breed*).

Make sure to include proper *matplotlib* visualizations (i.e. regression function) and tables to support your analysis. Also make sure your model is ***properly evaluated*** with the appropriate metrics and evaluation sets.

Your analysis must be a comparison between *Linear Regression and KNN Regression.*

4 Grading policy

This homework doesn't have any automated testing, or any other *strict* criteria of evaluation. As such, you will be graded based on the thoroughness of your analysis, and the presentation of results according to the following criteria.

The submitted code is sufficient for us to reproduce the results described in your report.	**5 points**
The project uses NumPy arrays and Pandas DataFrames where appropriate, rather than Python lists and dictionaries. Where possible, vectorized operations and built-in functions are used instead of loops.	**5 points**
The report documents any changes that were made to clean the data, such as normalization, handling missing values, etc.	**5 points**
You should state and describe the problem you are trying to solve. Describe the data (type of data, summary statistics), the desired output, and the proposed model.	**5 points**
You should describe any features transformation you make. You need to experiment with creating new features out of the available feature set. (One hot encoding, polynomial features, etc.)	**10 points**
Visualizations made in the report depict the data in an appropriate manner, which allows plots to be readily interpreted. Include both exploratory plots and plots to visualize the behavior of the model.	**10 points**
You should experiment with multiple values for hyperparameters. Describe the strategy used to find the best values.	**10 points**
Compare all models with all the hyperparameters, and select the best model for the classification problem. Justify your decision.	**15 points**
Compare all models with all the hyperparameters, and select the best model for the regression problem. Justify your decision.	**15 points**
Your decisions must be supported by arguments (i.e. table with hyper-parameter values, metrics used, other explanations etc.)	**20 points**
Bonus points. For a thorough analysis that goes beyond the scope of the course.	**15 points**
Total	**115 points**

5 Other Considerations

- Make sure to properly search for *hyper-parameters* for each of the models you train (e.g. k for *KNN*, regularization parameters etc.).
- *Longevity* and *Dog Breed* are the labels you need to predict. **Don't** use any of them as features.
- Some dogs have missing attributes. It's your decision how to deal with them.
- Each student will receive his own personal dataset. This means that the parameters you find will likely be unique and optimal for **your own** dataset.
- We also have a separate test set on which we will evaluate your models.
- **All code must be written in Python v3.6** (or greater).
- **Collaboration** is allowed and encouraged regarding techniques used, model details (how it works, how to tune it), descriptive statistics, etc. but keep in mind that the report and the code submitted has to be your own work.