# Introduction to Data Science Homework 10

**Title of Project:** Project C9. Kaggle – Dog Breed Classifier

**Team Members**: Ali Zeynalli, Paola Avalos Conchas, Iryna Hurova

Repository link:
https://github.com/PaoAvalos/DogBreedClassifier_UT

## Task 2: Business understanding (1 point)

### Identifying our business goals

- *Background*:

Sometimes you are walking on the street and see a beautiful dog, but you are too shy to ask the owner what the breed of their dog is. With our algorithm you can take a sneak picture of the dog and find out its breed without communicating/disturbing the other people, including the owner of the dog.

- *Business goals*:

Create an algorithm which can differentiate between the dog breeds.

- *Business success criteria:*

Good grade for the project.

### Assessing situation

- *Inventory of resources*:

For this project we will use tree laptops (HP ProBook 430 G7, MacBook Pro). For creating an algorithm of dog breed classification, we will use the web-based interactive computing platform Jupiter Notebook. The data that will be used is taken from the open-source KAGGLE competition called Dog Breed Identification.

- *Requirements, assumptions, and constraints*:

For successful completion of this project, we need to make an poster before the 12th of December and then present it on the poster session on 15th of December. Requirements for acceptable finished work are finished poster with description of the project and analysis of the results obtained while testing a trained model, poster also needs to contain multiple graphs that represent the analyzed data. In terms of the algorithm itself, it should predict the dog breed from the picture as accurately as possible (more than 90% accuracy required).

- *Risks and contingencies:*

Delay with the finishing the project due to other courses deadlines, tests and exams.

- *Terminology:*

We don’t have any specific terms.

- *Costs and benefits:*

No money was spent or gained by this project.

Defining our data-mining goals

- *Data-mining goals:*

For the completion of the project, we will need to test different models with various criteria to select the best one for our goal. The data preparation will also be a crucial part in model training, since we use the classification method based on image input, we will need to try out different image processing techniques to in order to see which conditions will lead to the best accuracy and overall best execution time.

- *Data-mining success criteria:*

Previously mentioned “best accuracy and execution time” are the two main criteria by which we will measure the success of the ongoing project. Expected accuracy set to be more than 90% to claim a model to be good enough for our task.

## Task 2: Data understanding (2 point)

The data-understanding phase includes four parts.

1. **Gathering data**

- *Outline data requirements:*

Image classification models usually need images and corresponding labels in order to train the model. The dataset provided consist of images of dogs in the dataset in “jpg” format, which is very common compressed format for digital images. The labels are saved in the csv file, called “labels.csv”, which contains of two columns: “id” and “breed”. The “id” is the unique name of the image in the train/test data and the “breed” is a corresponding label to that the image. The format of data for the current project is “jpg” (images), and “csv” (labels).

- *Verify data availability:*

The data is publicly available and is gathered from the official Kaggle website, under the competition "Dog Breed Identification". We are provided with training set and a test set of images of dogs. Each image has a unique id and filename. There are 120 different breeds of dogs. Moreover, to mention that the size of dataset is 750.43 MB.

- *Define selection criteria:*

As mentioned before, the source of the data is the from the official website of the Kaggle, which provides two folders "train" and "test", each consisting around 10k files, where each file is a "jpg" format image with unique id as a filename, also two "csv" files called "labels.csv" and "sample_submission.csv". There are 120 breeds of dogs, therefore 120 different labels. The "lables.csv" contains the same number as the number of files in train dataset, since the model will learn using train images and labels and predict using test images.

Our dataset for the project is less than around 750 MB, which is less than a GB, there will be no problem training and predicting the models in our personal computers' CPUs, unless some of us would like to get access to GPU's if needed, but for this dataset is not required. The author of this report for this current task has imported data, visualized the images, and has read the labels.csv files, just in case, to check the ability to read the data, and that the data sources are not damaged. Datasets dis compatible with data-mining platform and do not pose any imperfections or limits to start getting processed.

## 2. Describing data

The data source, including the dataset and datatypes are compatible with the requirements needed for the realization of this project, which leads to the achieving the goals of the business. The classification model requires images and labels, both are included in the dataset gathered from the Kaggle competition and are in suitable data formats to meet the data-mining goals. All the required fields are filled, and number of cases are suitable to achieve the reliable results of the performance.

## 3. Exploring data

Since, the goal of the project is to have a classification model with best performance on dog breeds, it is important to have balanced data for all the breeds (classes). Having checked labels and visualized the images, there is no sing of data quality problems was spotted. And the data is ready for further preprocessing stage.

## 4. Verifying data quality

As mentioned before, the data exists, it is available data source, has enough number of cases and fields required to reach the goal of the project. The quality of data is good enough to move to the next stage of the Data Processing. Later, in next stages, if there are any problems that pomp out, regarding the dataset, we can apply different techniques of data augmentation as a part of preprocessing.

**Task 4: Planning of project (0.5 points)**

**Detailed plan of the project with a list of tasks.**

| Task | Details | Approximate time | Team member |
|---|---|---|---|
| Organizing the data | Having a good organization for data as well as splitting into train and test groups | 5.5hrs | Iryna |
| Image processing and sorting. | Since we will be working with vastly different images it's important to process them. As well as sorting the size, rotation or other factors to standardize them. | 9 hrs | Paola |
| Filtering the images | We need to use filtering and equalizations to make sure the images can be used. | 6hrs | Ali |
| Build the model | Looking into how to use different models in order to find the correct one and implement it accordingly. | 21 hrs | Joint work |
| Optimize parameters of the model | We also need to define the hyper parameters to use, and find a way to look for the most efficient ones. | 7hrs | Iryna |
| Training the model. | Training and optimizing the model | 4hrs | Ali |
| Evaluate the model | Check the accuracy of the model and find ways to improve it. | 8hrs | Iryna |
| Creating graphs for the poster | | 3.5hrs | Ali |
| Creating the poster | | 4hrs | Paola |

| | | | |
|---|---|---|---|
| design | | | |
| creating an understandable README file. | In order for others to be able to replicate our work we need to create instructions in the readme of the repository so that others can use our work | 7 hrs | Paola |
| creating understandable and replicable code. | Polishing the code, making sure variables and methods can be understood by others that read our project and cleaning up the code to make it as readable as possible. | 6hrs | Ali |

- List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

So far we plan to use these tools, however some might be added in the future depending on the challenges that will arise.

For data handling:
os, numpy, pandas, cv2 (cv2 also for image processing and filtering)

For Visualization:
IPython.display (to look at the images and how they changed after the image processing) , seaborn, matplotlib (we will use this for making accuracy plots and understanding the numbers better)

For model building and training:
keras, BatchNormalization, Input, Flatten, MaxPooling2D, Lambda, UpSampling2D, Concatenate, Adam (keras optimizer), InceptionResNetV2, ImageDataGenerator. Keras is a very good tool for this case so we will be relying on it a lot.We saw that this tool is very used in other kaggle competitions alongside the TensorFlow environment. We will also use this backend.