## Project Structure

1. **Problem Definition**
   1. We will try to predict the price of the vehicle according to the following features: model, year of manufacture, gearbox, mileage, engine type (gasoline, hybrid, diesel and other), tax, engine size and fuel consumption by distance. We will try to build the new popular car price list for car lots.
   2. We would like to measure the average total error of vehicle prices by models. The goal is to reach an accuracy of up to 1000 [euros].
   3. We have the most necessary features, we would be happy to receive the type of vehicle if it is private, a taxi and in addition if it is a hatchback or sedan, etc. Although we do not have these features, our existing features are definitely enough to estimate the prices of the models.
   4. The proposed algorithm will be a website designed for car lots in order to estimate the purchase and sale prices of the models in their possession. Thus, it should be able to handle 500 entries/hour.
2. **Data Exploration**
   1. We have presented the data with the most information to understand what we deal with.
   2. The price has a strong correlation with class, engine size and some to age of the car. Mileage has strong correlation with age of the car, Mileage has strong correlation with tax and some more.
   3. The data looks dense without the outliers.. We believe that we will achieve a nice score.
3. **Data Pre Processing**
   1. We checked the integrity of the dataset and searched for incomplete and duplicated entries.
   2. Then, we continue to clean the data from outliers by using sigma equals 3 and remove any small groups under 30 entries.
4. **Feature Engineering**
   1. The column of year we translated to the age of the car. We added another column that describes the class of the car model.
   2. The price has a strong correlation with engine size and some to age of the car. Furthermore, age\_car has a strong correlation to mileage and tax. Although, engine size hasn't any correlation but the price.
5. **Regression**   
   1. We used linear regression model, tree model and knn models. We used all the methods to achieve the best score and least RMSE.
   2. We split the whole data to 70% learning data and 30% is test data. We used cross validation for tree models with max and min leaves.
   3. The linear regression model has scored 93.79 while the linear regression model has scored for the aygo model the score was 85.78.   
      We have created 4 hyperparameters: mileage per age\_car, mpg per group type of the car, engine size per car model (each model is sold with different engine size), engine size per the age of the car.