Exercise 2 : Applied Artificial Intelligence – ES Minor – Electrical Engineering

Lab2: A quick overview of machine learning project.

0: Read the rest of chapter 2 (section 5-8) End-to-End Machine Learning Project of the book [1].

The end goal of this lab is to train a linear regression model to make an insurance cost prediction. Write the necessary code and answer the questions below.

1. Use the CSV file from Blackboard **Med_insurance.csv [2]** from last week**.** It contains data of medical information and insurance cost. It contains 1338 rows of data with columns: age, gender, BMI, children, smoker, region, insurance charges. Read this csv file using pandas library into a variable called ***insurance_data***

2. Convert a categorical variable of your choice into dummy/indicator variables using the pandas function pandas.get_dummies() and combine the result with the numerical columns of the insurance_data

3. Create a test set which is 20 % of the whole data set using a pure random sampling approach.
   Question: Why do you need a test set when training a model?

Preparing the data for Machine Learning algorithms
In the upcoming steps you will prepare the data that will be used to train a machine learning model

1. If you found missing values in the data add the missing entries for the respective column(s) using the imputer transform (including only numerical attributes) in Scikit SimpleImputer class. Use "*median*" as strategy. Make sure to train the imputer only on the training set.

2. Perform feature scaling on all numerical attributes using Scikit transform *StandardScaler*. Again, fit the scaler on the training data only.
   Question: Explain what problem or problems the feature scaling resolves.

Save the prepocessed data into a variable *insurance_data_prepared.*

Training a model
In the following steps you will select and train a machine learning model.

4. Now your data is ready to be used in training a machine learning model. Use *it* to train a Linear Regression model that can predict insurance charges. See the *Training and Evaluation on the Training Set* section at page 72 in the book [1].
   Question: What is the difference between supervised and unsupervised learning? Is a Linear Regression supervised or unsupervised?
   Question: What is the difference between regression and classification?
5. Test your trained model with the test data and find out the *root mean squared error* and *mean absolute error*.