h_da

HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

ML Classification Project

Lecturer: Meghan Kane

Subject: Machine Learning

Author: Susanne Clara, 755769

Sommersemester 2020

28.08.2020

Content

1. Idea & Motivation

For my 2nd Machine Learning-project I wanted to learn more about the programming part as well as working with datasets and the programming language *python*. Also, I wanted to consolidate my understanding and the knowledge I gained in the last course by applying my knowledge to a programming project.

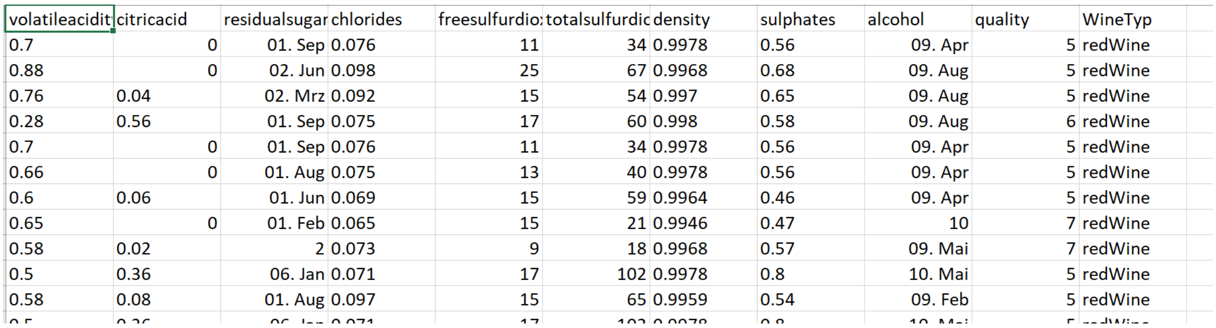
2. 4 Stages of ML

Framing Problem: Since the goal was to learn how to work with datasets and TensorFlow the problem was to find a good and useable dataset and decide what feature would be trainable and what to predict. My first try was to predict the quality rate of wine. Then I had the idea to add a wine-Type feature manually to the dataset (red- or white Wine) and predict this with a model. Before I found the wine dataset I tried several other data sources which I couldn't use because eather the bad quality of the dataset or the lack of meaningful features and values in the data.

Data Prep: Data Preparation was very useful in this project. I used Microsoft Excel to change the data into a format that fits for TensorFolw otherwise it wouldn't have worked to load the data into the program and devide it into different columns.

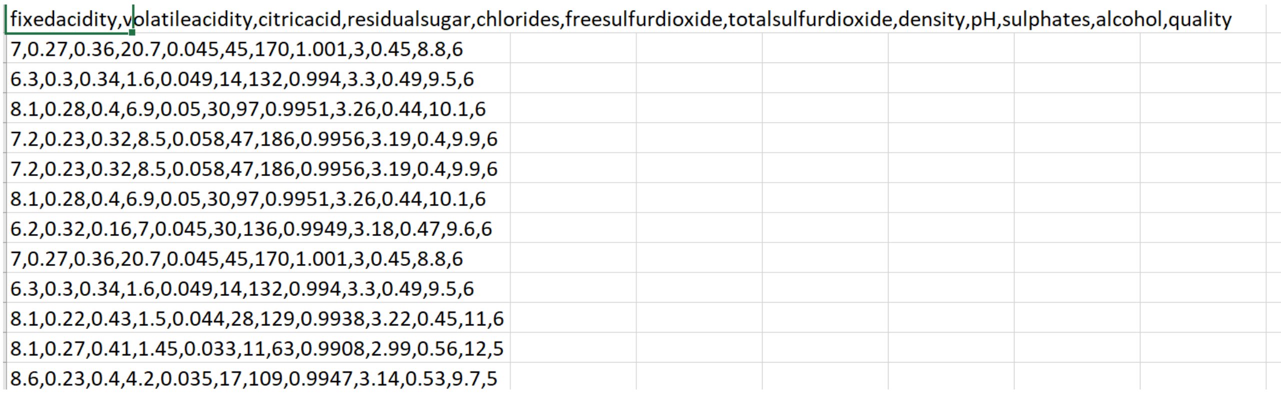
This is how it originally looked like:

Some numbers were automatically converted into dates by excel. Which would have been a critical error without Data preparation.



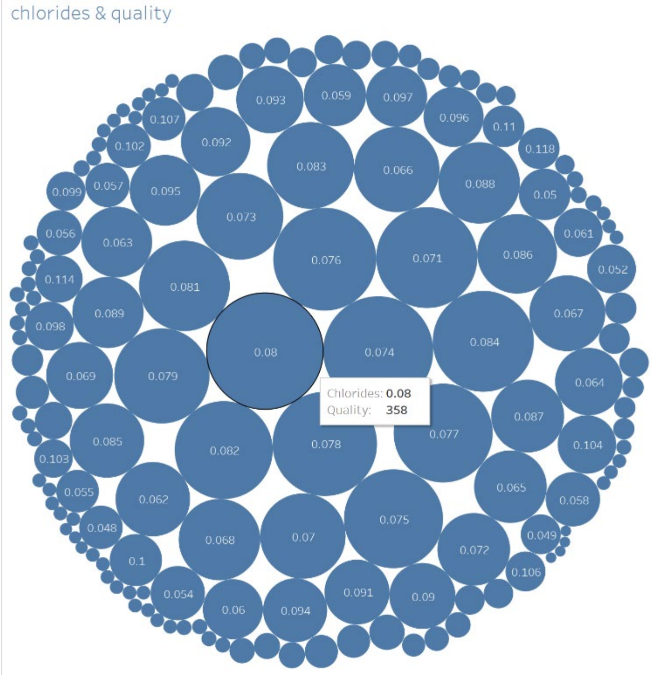
volatileacidit	citricacid	residualsugar	chlorides	freesulfurdio:	totalsulfurdic	density	sulphates	alcohol	quality	WineTyp
0.7	0	01. Sep	0.076	11	34	0.9978	0.56	09. Apr	5	redWine
0.88	0	02. Jun	0.098	25	67	0.9968	0.68	09. Aug	5	redWine
0.76	0.04	02. Mrz	0.092	15	54	0.997	0.65	09. Aug	5	redWine
0.28	0.56	01. Sep	0.075	17	60	0.998	0.58	09. Aug	6	redWine
0.7	0	01. Sep	0.076	11	34	0.9978	0.56	09. Apr	5	redWine
0.66	0	01. Aug	0.075	13	40	0.9978	0.56	09. Apr	5	redWine
0.6	0.06	01. Jun	0.069	15	59	0.9964	0.46	09. Apr	5	redWine
0.65	0	01. Feb	0.065	15	21	0.9946	0.47	10	7	redWine
0.58	0.02	2	0.073	9	18	0.9968	0.57	09. Mai	7	redWine
0.5	0.36	06. Jan	0.071	17	102	0.9978	0.8	10. Mai	5	redWine
0.58	0.08	01. Aug	0.097	15	65	0.9959	0.54	09. Feb	5	redWine

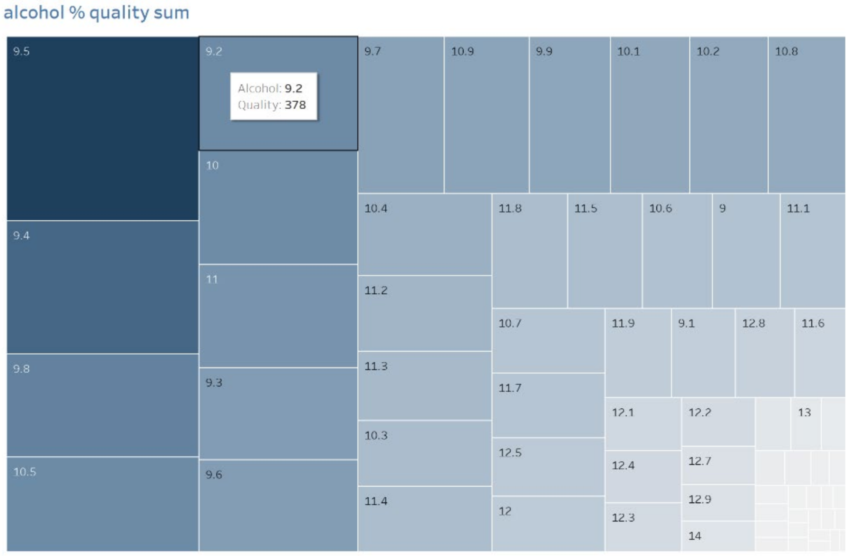
This is the useable outcome of my adjustments in excel:

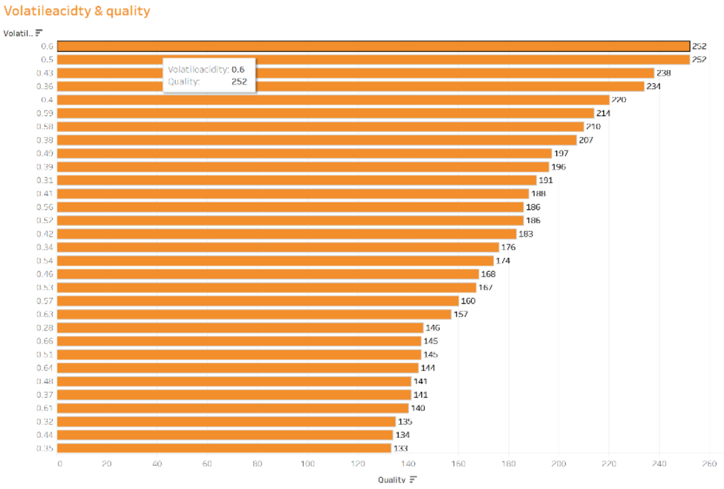


```
fixedacidity,volatileacidity,citricacid,residualsugar,chlorides,freesulfurdioxide,totalsulfurdioxide,density,pH,sulphates,alcohol,quality
7,0.27,0.36,20.7,0.045,45,170,1.001,3,0.45,8.8,6
6.3,0.3,0.34,1.6,0.049,14,132,0.994,3.3,0.49,9.5,6
8.1,0.28,0.4,6.9,0.05,30,97,0.9951,3.26,0.44,10.1,6
7.2,0.23,0.32,8.5,0.058,47,186,0.9956,3.19,0.4,9.9,6
7.2,0.23,0.32,8.5,0.058,47,186,0.9956,3.19,0.4,9.9,6
8.1,0.28,0.4,6.9,0.05,30,97,0.9951,3.26,0.44,10.1,6
6.2,0.32,0.16,7,0.045,30,136,0.9949,3.18,0.47,9.6,6
7,0.27,0.36,20.7,0.045,45,170,1.001,3,0.45,8.8,6
6.3,0.3,0.34,1.6,0.049,14,132,0.994,3.3,0.49,9.5,6
8.1,0.22,0.43,1.5,0.044,28,129,0.9938,3.22,0.45,11,6
8.1,0.27,0.41,1.45,0.033,11,63,0.9908,2.99,0.56,12,5
8.6,0.23,0.4,4.2,0.035,17,109,0.9947,3.14,0.53,9.7,5
```

Since the right format is not the only thing that is important before training and writing a Machine Learning mode, but also to understand the data that is used. Finding out what parts of the data are important or significant for the task is a crucial part of the data preparation step. Therefor it sometimes can be useful to apply data visualization tools to your data. This could show if the data is clean and which features you want to train the model on. In this project I used the tool ‘tableau’ for the data analysis and visualization. In the following you can see a couple of visual outcomes of this step:







The biggest learning in this step was not to get confused by the summation of the quality number, which could mislead, since the sum doesn't show the ones with the highest quality. What sounds confusing is a really important fact to know when working with this data. Since in this case the program sums up all wines with for example 0.6% acidity. But that doesn't mean the biggest sum, has the best quality. Sometimes these visualizations can be misleading. But this step still helped getting to know my data.

Training the Model

For this project I used TensorFlow and Keras to build a model that could apply a pattern classification to the data I feed it with. I learned about what a sequential model is and that has exactly one input and one output layer. I also learned about normalization of the prepared data.

Predictions

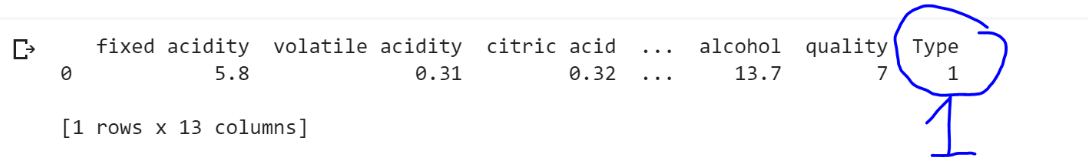
In the documentation I found two methods of feeding new data into the model.

1. model.predict()
2. model.evaluate()

With the first method you can take a single input and predict an output. The second method lets you test your model and its accuracy with a bigger dataset, that the model hasn't been trained with before to test if the models makes accurate prediction in general. I tried both methods and this was the outcome:

Model.Predict():

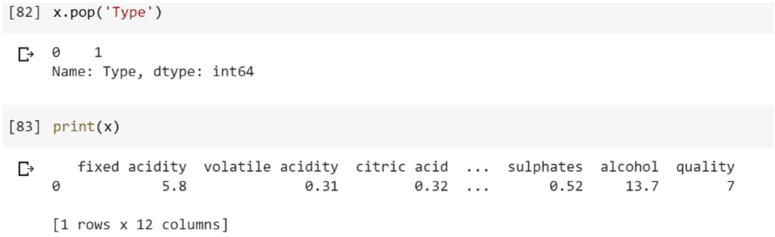
Fist upload the new Data:



```
   fixed acidity  volatile acidity  citric acid  ...  alcohol  quality  Type
0            5.8              0.31         0.32  ...     13.7        7     1

[1 rows x 13 columns]
```

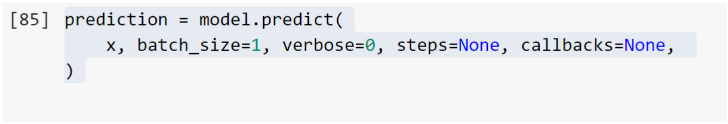
1

Then delete the type feature:

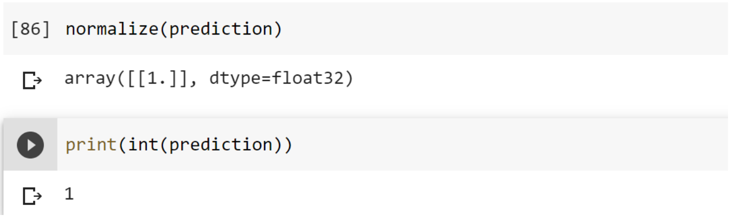


```
[82] x.pop('Type')

     0    1
     Name: Type, dtype: int64

[83] print(x)

        fixed acidity  volatile acidity  citric acid  ...  sulphates  alcohol  quality
     0            5.8              0.31         0.32  ...       0.52     13.7        7

     [1 rows x 12 columns]
```

Prediction: Right prediction of type: 1

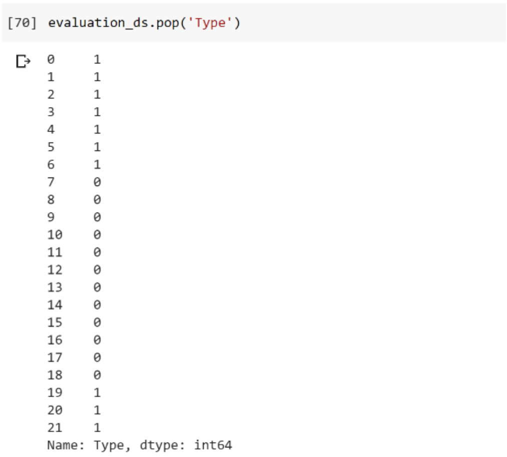


```
[85] prediction = model.predict(
         x, batch_size=1, verbose=0, steps=None, callbacks=None,
     )
```



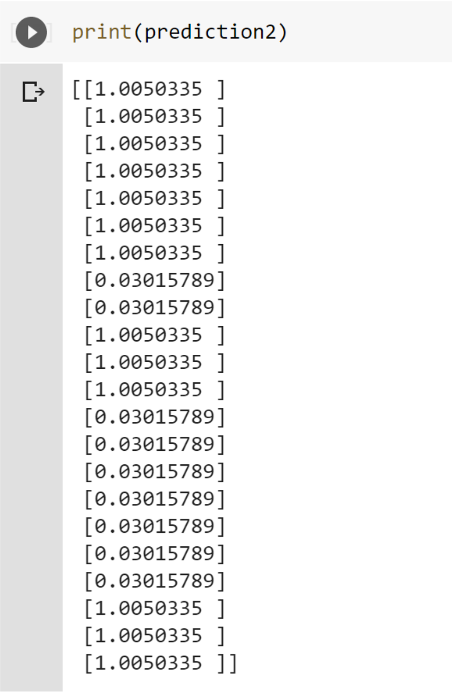
```
[86] normalize(prediction)

     array([[1.]], dtype=float32)

     print(int(prediction))

     1
```

Model.Evaluate():



← Input, 22 rows, new data

```
[70] evaluation_ds.pop('Type')

    0     1
    1     1
    2     1
    3     1
    4     1
    5     1
    6     1
    7     0
    8     0
    9     0
    10    0
    11    0
    12    0
    13    0
    14    0
    15    0
    16    0
    17    0
    18    0
    19    1
    20    1
    21    1
    Name: Type, dtype: int64
```



```
print(prediction2)
```

```
[[1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [0.03015789]
 [0.03015789]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]
 [0.03015789]
 [0.03015789]
 [0.03015789]
 [0.03015789]
 [0.03015789]
 [0.03015789]
 [0.03015789]
 [1.0050335 ]
 [1.0050335 ]
 [1.0050335 ]]
```

← Prediction: with seemingly 3 wrong prediction

3/22 wrong = 16,63% wrong prediction

19/22 right =   86,2637% right prediction

3. Learnings

Dense Layers:

The learning with dense layers was, that they make the network a deep learning network. With each layer you add more neuron layers and make the network deeper with more calculation and depth.

Model structure:

I learned how to structure a code for a ML project. The biggest learning was how to use Dense Layers and prediction functions and how to use activation functions.

Python + np + pandas + keras + matlibplot:

Working with python for the first time I found out that there are some standard libraries that help when working with data and Machine Learning. Such as: Numpy, Pandas, Keras and matplotlib.

Evaluation and Prediction:

softmax function for classification problems: Softmax function, a wonderful *activation function* that turns numbers aka logits into probabilities that sum to one. Softmax function outputs a vector that represents the probability distributions of a list of potential outcomes.

Running ML locally on my CPU

Fortunately, I was able to figure out how to run TensorFlow locally on my CPU with the Jupyter Notebook and the Anaconda GUI.

Outcome & Problems

The Outcome of this project are two models. One for predicting the wine type (red or white wine) based on the chemical Ingredients given by the dataset, and one model for predicting the wine quality based on the same data.

The ‘type prediction’ was right and 100% sure in all the prediction. The quality predicting fairly close to the test data bit never really good. So in the future to fix this problem, it would be necessary to redo all the ML Steps and analyse if the error comes from the dataset or the tf.model structure.

4. Future Goals

Website Interface to feed Data and control the model through an Interface

One big goal that I will try to accomplish after this course finished is to include the model as a backend into a website, where it should be possible to type in the ingrediences of a wine a predict the quality or the type.

Explorer Tensor Flow Lite

From listening to Lex Fridmanns Podcast I learned more about Tensor Flow Lite. It sounds interesting to me so in my spare time it will probably be very fun and interesting to check out the possibilities of this framework for mobile and edge devices.