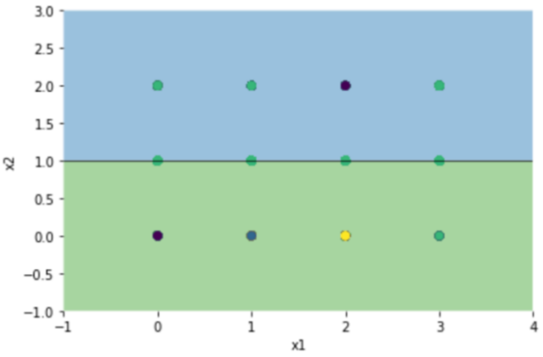
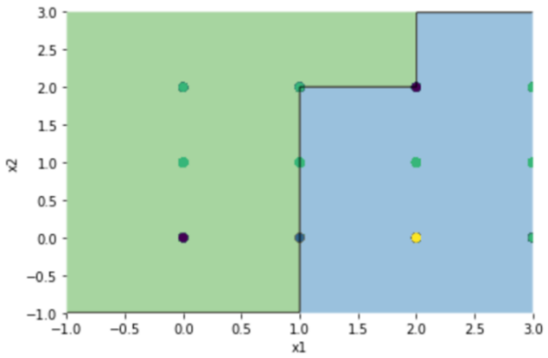
Cars

Logistic Regression

- **eta** of 0.01 yields the best results. (0,56 and 0,55 for training and testing data respectively). Increasing and decreasing it makes the final scores worth, 0,53 and 0,50)
- **iterations_num** of 100 achieves an unacceptable score of 0.45 for test data. A threshold of 1000 iterations is needed in order to achieve better results than a randomness (50/50). 10000 iterations for instance get 0,575 score.
- Using a higher **regularization term (C)** increases significantly the prediction scores. For example, with C of 100 we get a score of 0.63. Anything above or below this number would decrease the prediction scores.
- Comparing different features in the plotting yields different decision regions. One of such comparisons is for *feature_index=[2,3] VS feature_index=[3,4].*





Random Forest

- Using 100 trees yields a more comprehensive plot than anything smaller than that. Consequently, we can observe that it's somewhat computationally heavy.
- Sample size of 20 also seems to be performing the best. \
- min_leaf of 5 seems to be optimal.

Note: since there is no scoring implemented for the RF we can only observe the plot and its changes based on different parameters.

Conclusion:

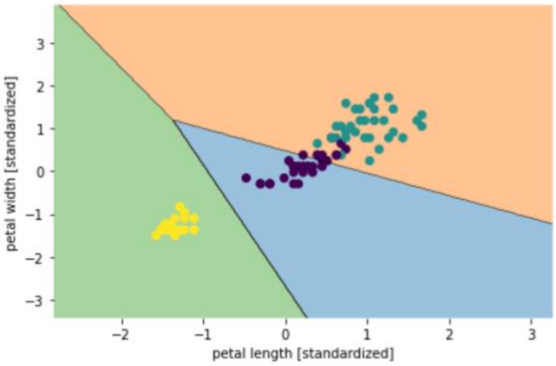
- We can observe RF yielding a more accurate plot, however since we don't know its scores it is hard to compare which one has better predictions.
- RF might be better to define a well-defined region.
- The overwhelming prevalence of *unacc* class makes the model harder to predict accurate classification.

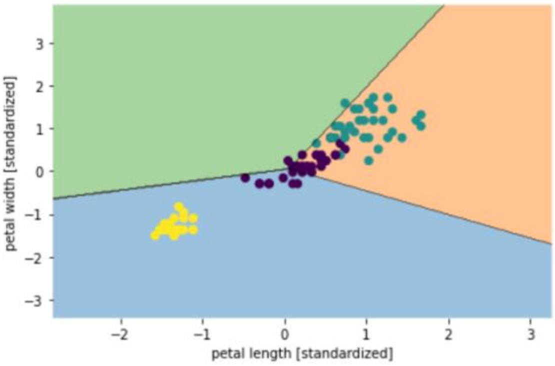
Iris

Logistic Regression

- Higher regularization terms distort the results a lot. C=10 for instance makes prediction below 0,5. From C of 1 to C of 5 we can observe that it stats to overfitting of C=4 with results of 8,5 (training) and 8,0 (test).

  With a number of iterations of 1000:
  C = 5 (on the left): yields a decently good result 0f 0,78 with a good separation of classes on the plot.
  C = 0.1 (on the right): yields a better result of 0,82, however the plot seems to capture less correct points for these features.



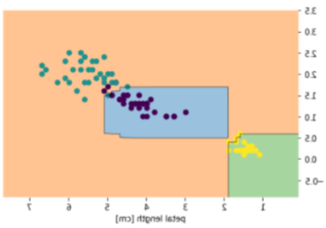


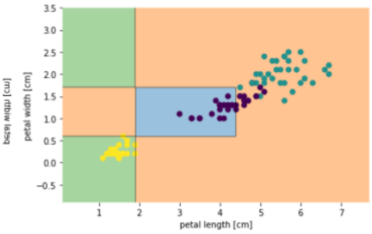
  However, the best result is actually achieved with iteration number of 10 and a C of 100 and an eta of 0.1 – 0,88 to 0,86. Such a drastic change of parameters complement each other the best.

- In order to further improve the model, change regularization kind - L1/L2.
- As well as improve the way the weights are updated
- An eta of 0,01 yields the best results. Anything more makes the scores close to 0,3. And anything smaller, provide the same scores although the plot captures the data much worse.

Random Forest

- Increasing sample size and decreasing the number of leaves seems to provide the best score and plot classification.
- Starting with 10 trees the model provides already good results.
- The score function implemented shouldn't be trusted for comparisons since it's very arbitrary.
- Sample size >50 overfits (observe how it performs with a 100).


petal length [cm]
petal width [cm]


petal width [cm]
petal length [cm]