DAT405/DIT406 Assignment 2 - Introduction to Data Science and Python

Artur Gasparyan Jessica Gilmsjoe

November 2021

Name	Time Spent
Artur	14 h
Jessica	15 h

Question 1

a)

The data cleaning started with checking for plausible NaN values in the relevant data for the model and no cleaning was required. The next step was to investigate potential outliers in the data. Detection of outliers started with plotting the data together with a regression line from a model fitted with linear regression which indicated a few outliers that were far away from the other points and the regression line. The residuals (see figure 1) whose absolute value was greater than 1.7 million were chosen as outliers, and these observations were removed from the data set. The outlier-cutoff was visually determined by observing the graph.

The cleaned data was fitted with a linear regression model and a scatter plot of the data together with a regression line is shown in Figure 2.

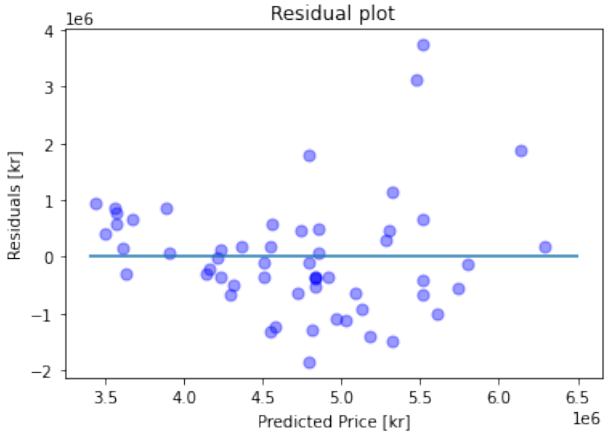

Figure 1: Scatter plot of residuals before cleaning.

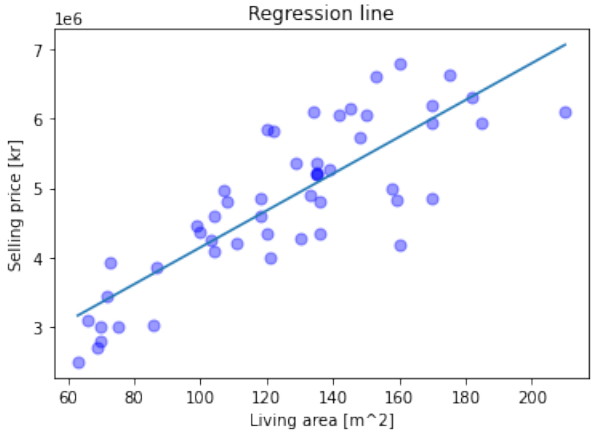

Figure 2: Scatter plot of living area and selling prices in Landvetter together with a regression line from the fitted linear regression model.

b)

The slope of the model is

$$26493.0947257$$

and the intercept is

$$1496944.97559781$$

.

c)

The predicted value of houses with areas 100 m^2, 150 m^2 and 200 m^2 obtained by the model is shown in Table 1.

Living area (m^2)	Predicted price (kr)
100	4146254
150	5470909
200	6795564

Table 1: Predicted selling prices of different living areas.

d)

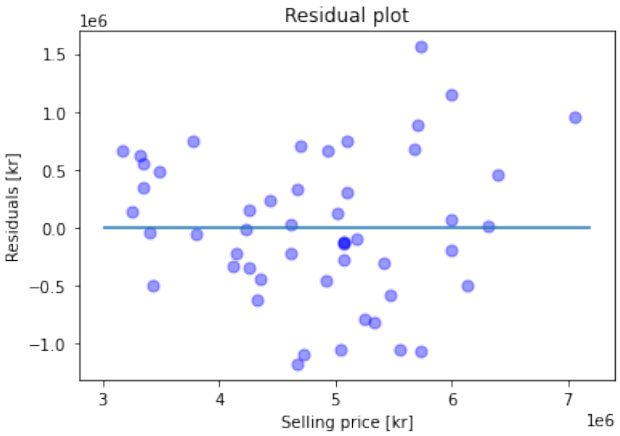

Figure 3: Residuals

e)

After removing the chosen outliers, the regression line fitted the data better and the residuals resulted in a more evenly distributed scatter around zero-line. There is a possibility to improve the model by removing more outliers but one might want to be careful not to lose potentially valuable information from the data. The residuals do not look particularly skewed, otherwise a transformation such as a log transformation of the data might be worth investigating. A regularised model could also be a way to improve the model.

Question 2

a)

The Iris data was divided into a training set including 75% of the data and a test including the rest of the data. Further, the data was fitted with a logistic regression model and the corresponding confusion matrix is shown in Figure 4.

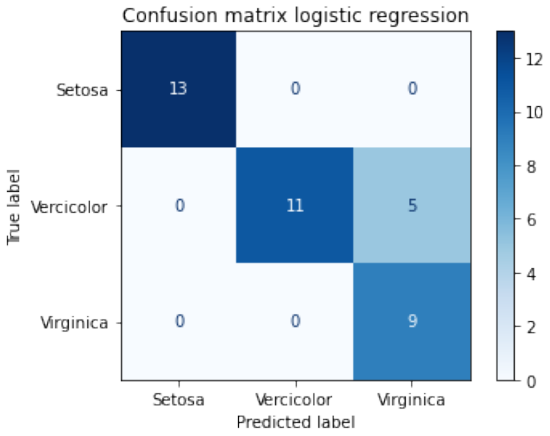

Figure 4: Confusion matrix of the logistic regression model.

b)

The Iris data was classified with KNN using the same proportions in the training set and test set as in previous task with different values of k neighbors with both uniform and distance based weights. For $k = 1$, there was only one misclassified sample and by increasing k, the predictions and evaluation metrics remained the same. There was no difference between uniform and distance based weights either. One possible reason for this could be the fact that the overlap is fairly small between classes, as seen in figure 5. Another reason could be that the data set is quite small, which leaves less opportunity for the data to intersect.

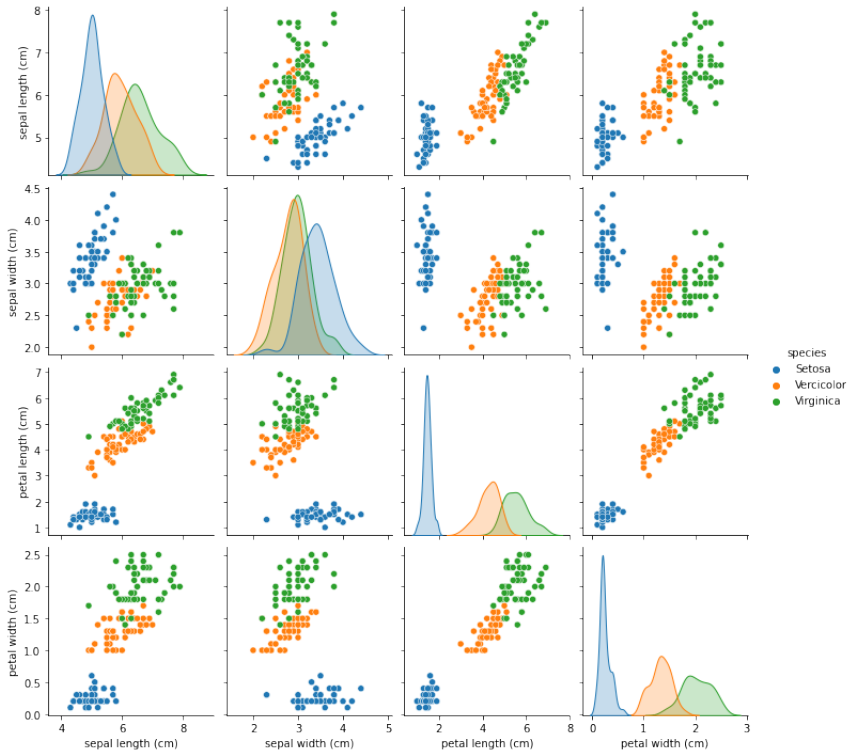

Figure 5: Pair plots matching the characteristics of the plants.

c)

Looking at figure 4, one can see that Vercicolor was incorrectly labeled as Virginica 5 times. Comparing that to the single mislabeling in the KNNs in figure 6 and 7, one can see that the accuracy greatly increases by switching to KNN. As previously discussed, the data is highly grouped, which is a prerequisite for increased high performance of KNN-classification. By looking at figure 5, one can see that the Vercicolor and Virginica classes blend together, which could be a reason for the increased frequency of mislabeling. These arguments are supported by the metrics seen in tables 2, 3, and 4.

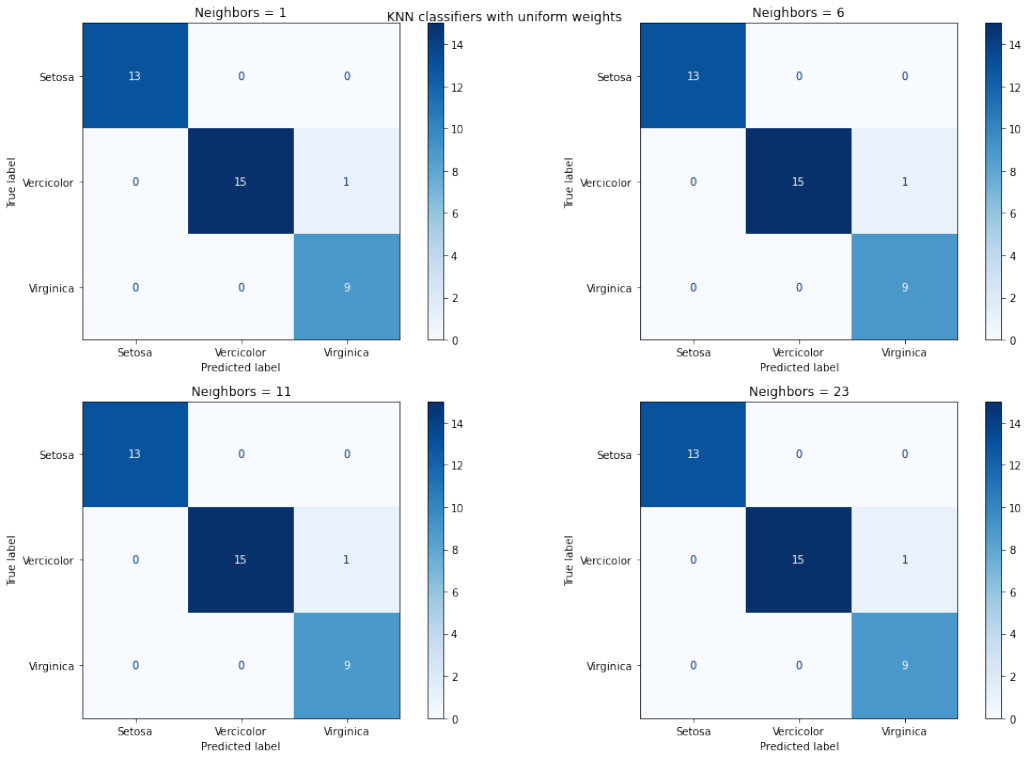

Figure 6: Confusion matrices of the KNN models for different values of k neighbors using uniform based weights .

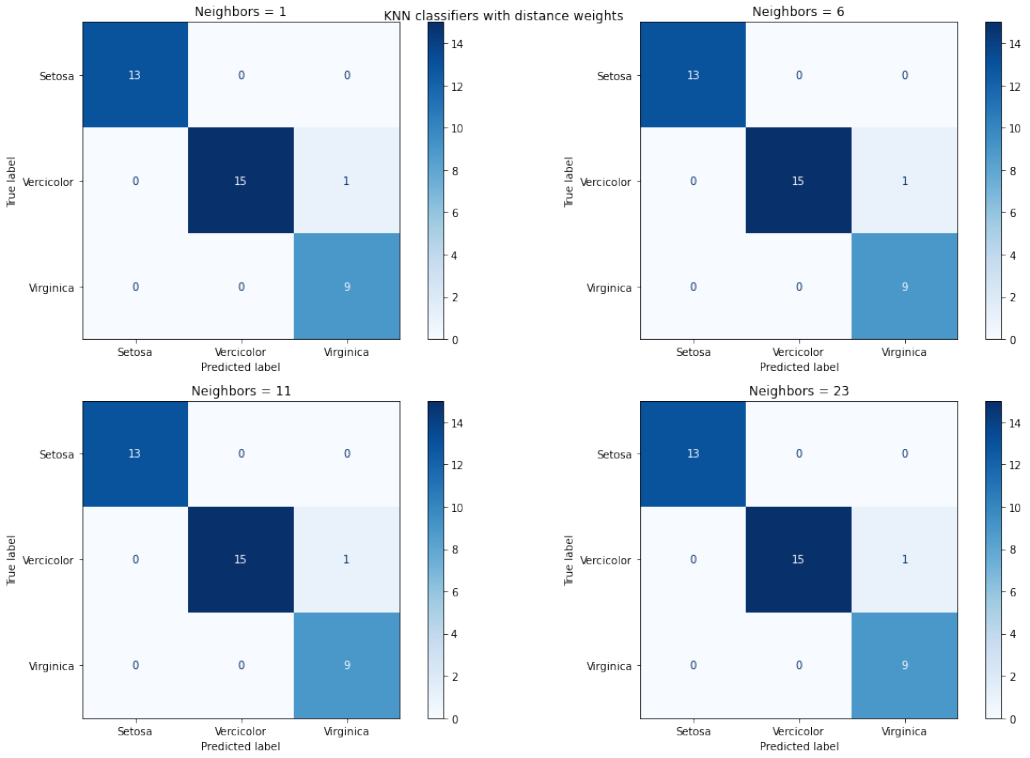

Figure 7: Confusion matrices of the KNN models for different values of k neighbors using distance based weights .

	Accuracy	Precision	F-score	Recall
Setosa	0.868421	1.000000	1.000000	1.0000
Vercicolor	0.868421	1.000000	0.814815	0.6875
Virginica	0.868421	0.642857	0.782609	1.0000

Table 2: Logistic regression metrics.

		Accuracy	Precision	F-score	Recall
1	**Setosa**	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
6	**Setosa**	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
11	**Setosa**	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
23	**Setosa**	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000

Table 3: KNN uniform based weights metrics.

		Accuracy	Precision	F-score	Recall
1	Setosa	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
6	Setosa	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
11	Setosa	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000
23	Setosa	0.973684	1.0	1.000000	1.0000
	Vercicolor	0.973684	1.0	0.967742	0.9375
	Virginica	0.973684	0.9	0.947368	1.0000

Table 4: KNN distance based weights metrics.

Question 3

The goal with the training itself is to find a model that can make predictions on unseen data with as high accuracy as possible. Without test data, it is difficult to know anything about the performance of the trained model on new data. The validation set can be used to do model selection by tuning parameters and later on evaluate the performance of the final model on the test data.