Assignment 2: Regression and classification

Laboratory Report
in DAT405 / DIT405 Introduction to data science and AI

Matilda Sjöblom Time spent: 6h
Simon Duchén Time spent: 6h

Group 7

November 18, 2020

1 Hemnet

a. Linear Regression Model
The linear regression model that fits the given data is: $y = 16.3x + 51.84$.
Where y is living area in square meters(m^2) and x is the price in millions Swedish crowns(M SEK). We did not data cleaning as no entries were NaN and the columns matched in length. It felt more true to reality to keep all the data points.

b. Slope value: $16.3 \cdot \frac{m^2}{MSEK}$, Interception: $51.84m^2$.

c.
$100m^2 = 16.3x + 51.84 => 48.16 = 16.3x => x = 2.954601$ M SEK

$150m^2 = 16.3x + 51.84 => 98.16 = 16.3x => x = 6.022086$ M SEK

$200m^2 = 16.3x + 51.84 => 148.16 = 16.3x => x = 9.089570$ M SEK

d. Residual plot

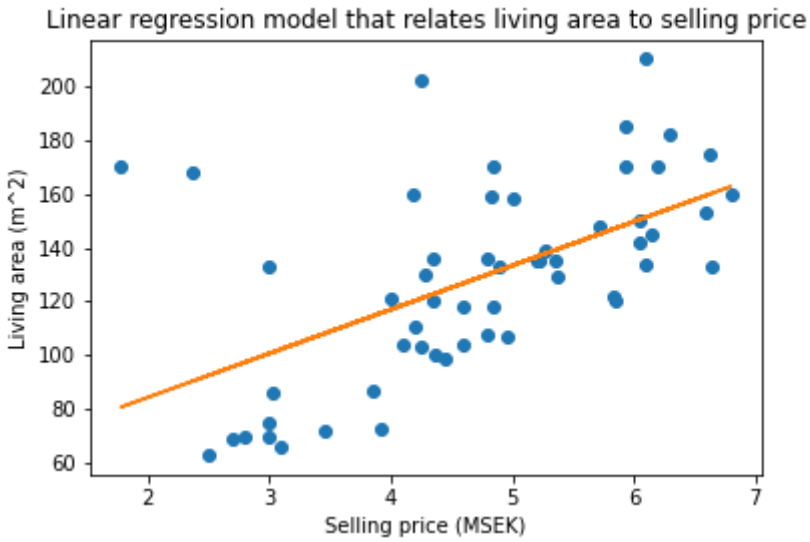


Figure 1: A residual plot showing the correlation between living area and selling prices and our linear model.

e. With almost any domain, the results of the analyses gets more interesting with more data points. In this set, there were only 56 data points, and with more points than that the regression could be more visible. Another thing that could improve the model is to delete the outliers, i.e. the four data points at the top of the plot. They are not so many, but are a long way from the model, so they are probably affecting the slope quite much.

2 SKLEARN

a. Below is the confusion matrix for the logistic regression. We chose the size of the training set to be 75% of the total size of the iris data set. The diagonal of the matrix is showing the True Positives. The accuracy was 97.37%.

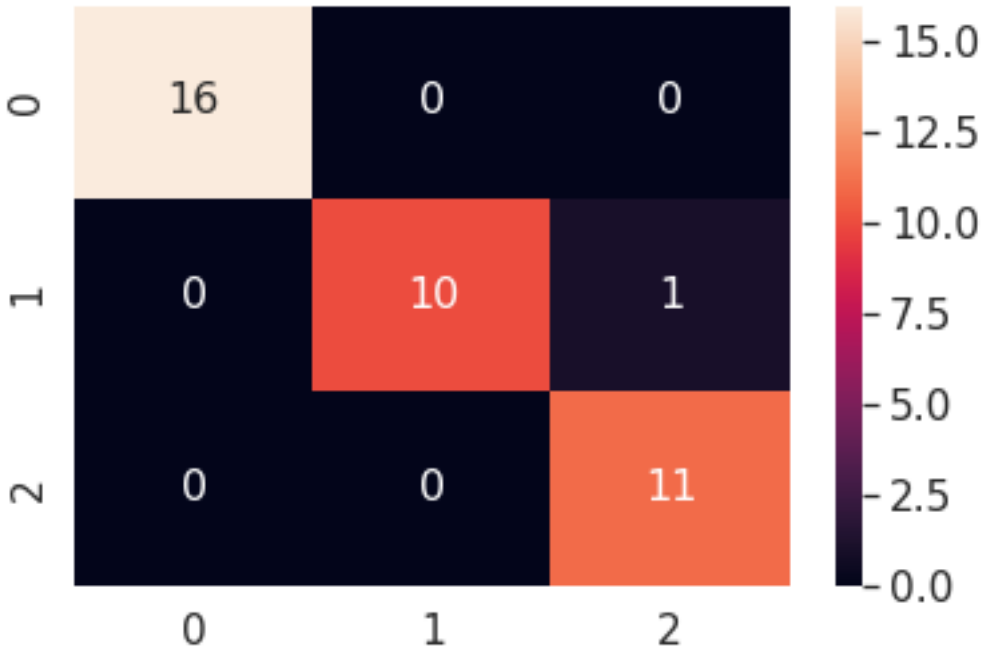


Figure 2: A confusion matrix that evaluates the logistic regression model on the iris dataset.

b. Below is two graphs showing the change of accuracy with different values of K. The first graph, fig 3, shows the change for distance based weights, and the second graph, fig 4, shows the change for uniform based weights.

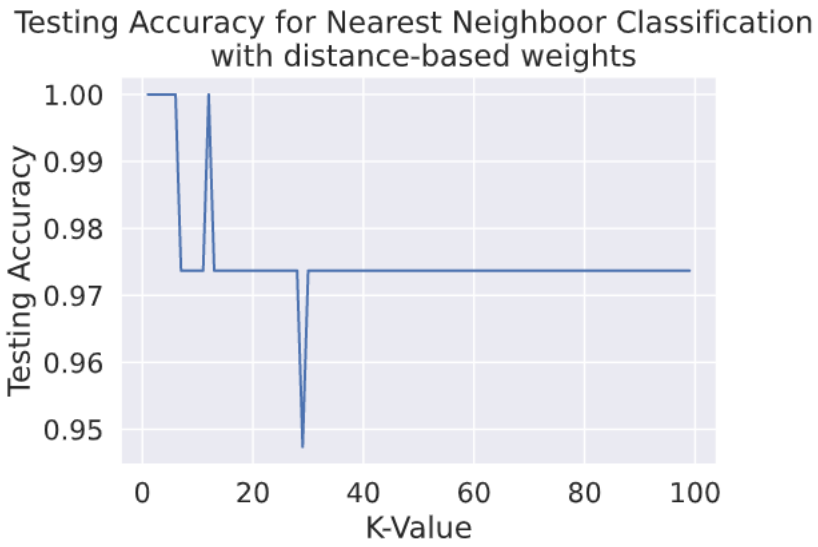


Figure 3: A graph showing the test accuracy with distance based weights for different values of K.

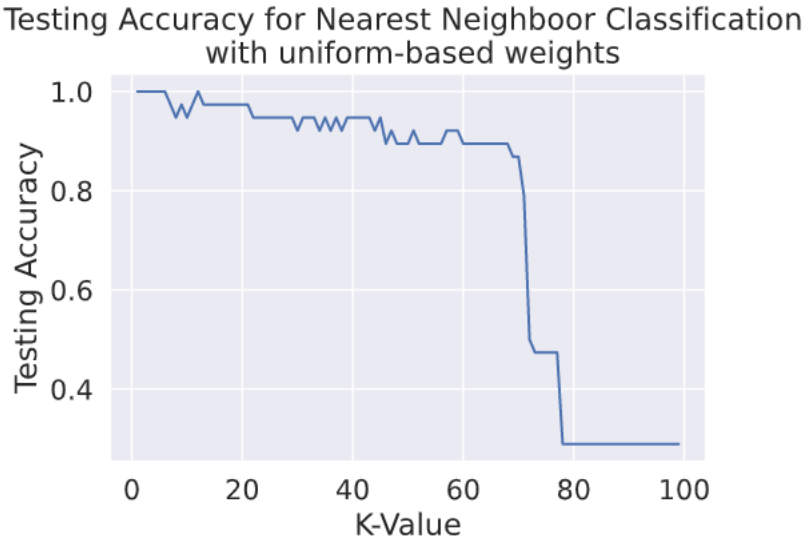
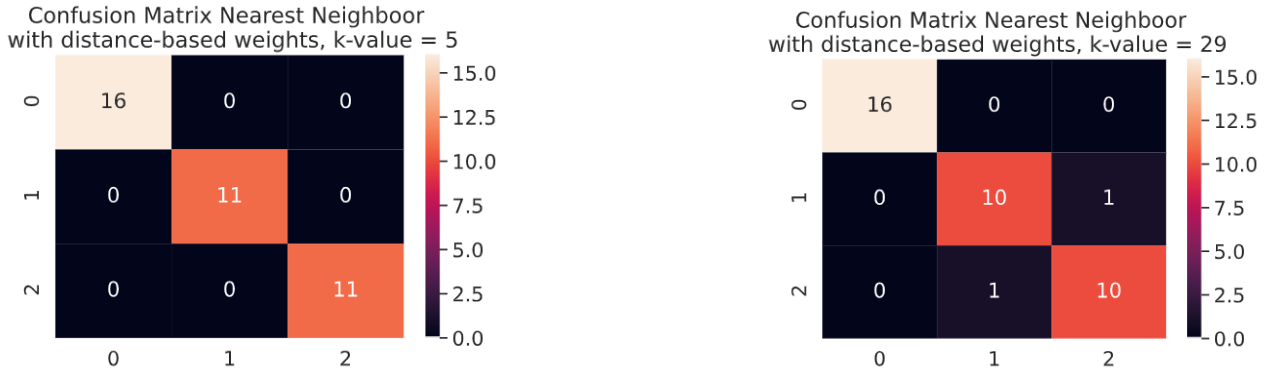


Figure 4: A graph showing the test accuracy with uniform based weights for different values of K.

As can be seen in the graphs above, a high value of K is never good for the accuracy. As there are three different classes, the accuracy should never go below around 33%, depending on the distribution of the test set. The uniform distribution weighs all the K data points equally, and therefore, if K is large, the class that has more data points will be more likely to grow. This will lead to more faulty classified data points the higher K is. With distance-based weights on the other hand, the data point that lies closest weights more than the data point that lies most far away, thus making the selection of class not only based on how many of the K nearest neighbours are of a certain class, but where they are placed.

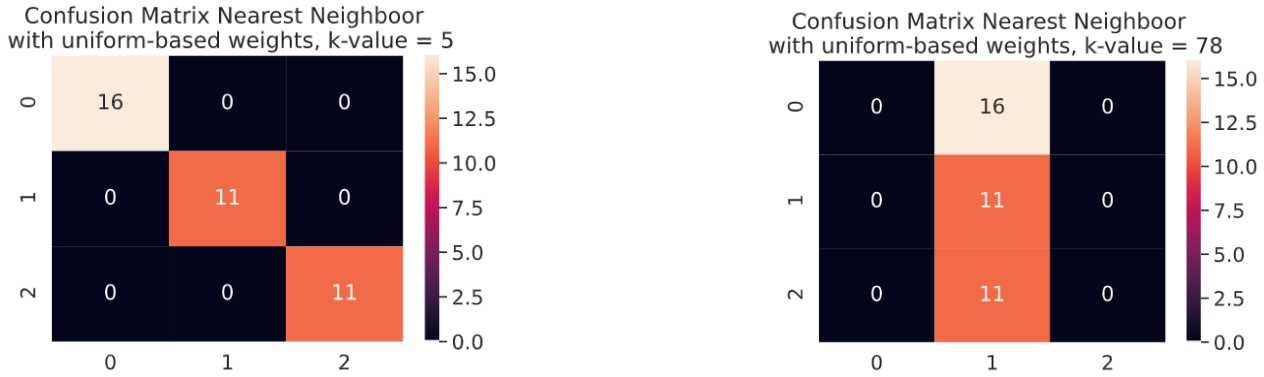
c. For the comparisons, we made four different confusion matrices for the KNN classifier. Two matrices with one of the best K's in terms of accuracy, fig 5a and fig 6a, and two with one of the worst K's, fig 5b and fig 6b. The confusion matrix for logistic regression can be seen in fig 2



(a) K-Value = 5

(b) K-Value = 29

Figure 5: Confusion matrices for nearest neighbour classification with distance-based weights



(a) K-Value = 5

(b) K-Value = 78

Figure 6: Confusion matrices for nearest neighbour classification with uniform-based weights

As can be seen when comparing the confusion matrices for the KNN classifier with the logistic regression, the KNN classifier performed slightly better. A reason for this could be that logistic regression, unlike the KNN classifier, does not work with non-linear data. Some of the data points in this set may be non-linear, and therefore the KNN classifier could perform better.

3 Different Sets of Data

The training of a model needs to done on a dataset. The model is shaped by this set, and learns to adjust itself to fit it. If you were to test the model on the same dataset as you trained it with, the model will be 100% accurate, as it was made to fit this data. Therefore, you cannot test the true accuracy of the model on the same dataset as you trained the model with. If you instead split your dataset in one training part and one testing part, you can test your model on data that is similar to the data it was trained for, but not exactly the same. This will give you a better idea of how well your model performed in terms of accuracy.