**DAT565**

**Introduction to data science and AI**

**Assignment 2**



***Keith Ahern***

***Alejandro Canals***

***Aaron McCann***

***1. 1 Hemnet Dataset***

*1. Find a linear regression model that relates the living area to the selling price. If in doing so, you performed any data cleaning step(s), describe what you did and explain why.*

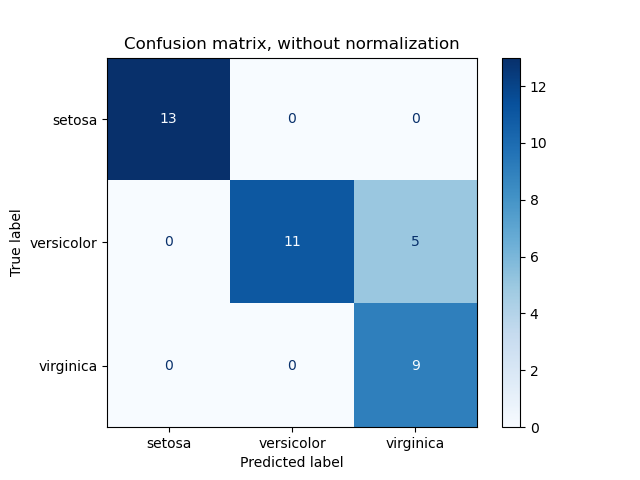
*2. What are the values of the slope and intercept of the regression line?*

*3. Use this model to predict the selling prices of houses which have living area 100 m2, 150 m2, and 200 m2.*

*4. Draw a residual plot. Discuss some potential strategies for improving the model.*

***1.2 Iris Dataset***

*1. Use a confusion matrix to evaluate the use of logistic regression to classify the Iris data set.*



From the confusion matrix above we can see that the logistic regression is very accurate at classifying the actual setosa and virginica categories. However, when predicting the versicolor category it sometimes mistook them for virginica. This can be seen by the 5 showing up in the square outside of the diagonal.

*2. Use k-nearest neighbours to classify the Iris data set with some different values for k, and with uniform and distance-based weights. What will happen when k grows larger for the different cases? Why?*

The following results were found when testing the accuracy of the various values of k and weightings:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Weights | K = 1 | K = 10 | K =25 | K = 50 | K = 100 |
| Uniform | 0.97777 | 0.97777 | 0.95555 | 0.91111 | 0.6 |
| Distance | 0.97777 | 0.97777 | 0.97777 | 0.97777 | 0.97777 |

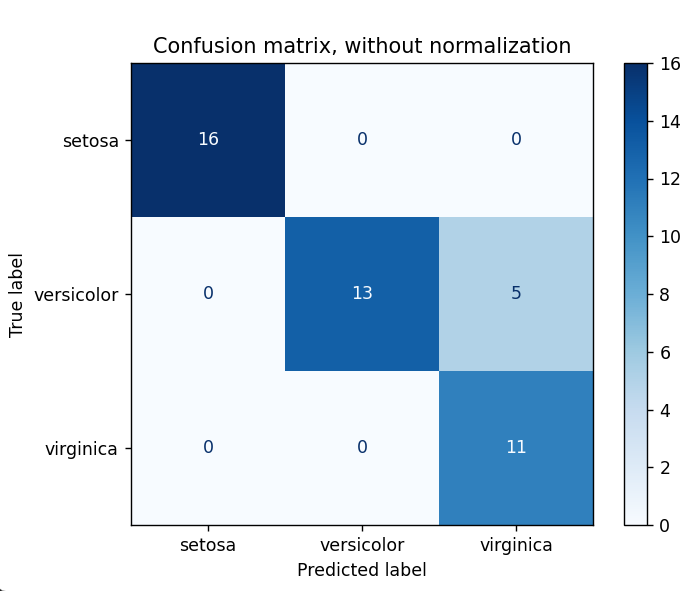
As seen from the table above the dataset was broad enough for little difference to be seen in the accuracy of the predictions until the k value reached a large value like 100.

When the k value reached 100 it significantly dropped down to 60% accuracy indicating that it would not be a very efficient value to select when trying to predict the correct category.

Distance weighting proved to be the most effective regardless of the value of k. This indicates that maybe the weighting placed on closer neighbours was significantly greater than those far away.

*3. Compare the classification models for the Iris data set that are generated by k-nearest neighbours (for the different settings from question 2) and by logistic regression. Calculate confusion matrices for these models and discuss the performance of the various models*

Logistic Regression



K Nearest Neighbours

|  |  |
| --- | --- |
| Uniform | Distance |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

As shown by the above figures. The k nearest neighbours method is more accurate than the logistical regression method, assuming either low values of k are used or the distance based weighting is used.

Logistical regression scored an accuracy of 88.9% while the best K nearest neighbour methods could score an accuracy of 97.8%, although it could go as low as 60& with uniform weighting and a k value of 100.