Assignment 1

av

Lars-Erik Bakkland Moi and Oskar Markussen

i

IKT450
Deep Neural Networks

Universitetet i Agder

Grimstad, August 2023

Contents

1 Task

In this task you are suppose to implement K-NN algorithms using only Python.

- Download the Pima Indian dataset: https://www.kaggle.com/kumargh/ pimaindiansdiabetescsv
- Implement K-NN classifier
- Predict the two classes
- Play with the hyper parameter K to find the best fit and evaluate the performance.

2 K-NN classifier

A K-NN classifier when used looks at the nearest neighbor and tries to uses those to predict what the given class is. In our assignment we implemented a K-NN classifier with the library "KNeighborsClassifier", we further created an algorithm to see the Accuracy, Recall, Precision and F1 Score of the validating data-set. With this we created an algorithm to test a range of k values, meaning how many neighbor we look at, to see which of the k values gave the most optimal output.

We tested a range from 1 neighbor to 50 to see which of the k values was the best fit, based on our results we found out that the k value with the best Accuracy was the value 8, which had a 0.77922078 Accuracy rating and was the top result. This result was also shared by another which was the k value 6. The k values that had the best Recall rating was the k value of 1, which had a rating of 0.58333333. The k value with the best Precision rating was the value 6, which had 0.68421053. Lastly the k value with the best F1 score was the value 8, with a rating of 0.61363636. We also plotted them to get a overview of every k in terms of Accuracy and F1 rating.

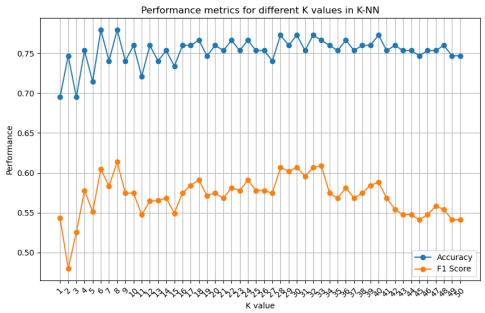

Figure 1: plot of accuracy and f1

From this incentive we can observe that both the k value of 6 and 8 had the highest among two performance evaluations, which was k value 6 had the highest in both Accuracy and Precision,

while k value 8 had the highest in both Accuracy and F1 score. While k value 1 had the highest value in Recall rating. For us to evaluate these result we first have to look at which performance evaluations is the most important in our Assignment.

Accuracy: How often is the model right overall? Recall: Out of the people who truly have diabetes, how many does the model catch? Precision: When the model says someone has diabetes, how often is it right? F1 Score: How well does the model balance its precision and recall?

Our Assignment involves identifying people with diabetes based on a set of factor, while we would argue that all performance evaluations are important in this aspect. We would believe the best evaluation of the model is to look at its F1 score based on the fact that both recall and precision is important to an issue such as identifying people with diabetes. It is important to catch the most people with diabetes and be right when it is called out and we will thereby say that the k value of 8 is the best fit for our project since its overall Accuracy and F1 Score is the best. We wanted to illustrate the Receiver Operating Characteristic(ROC), which is the True Positive rate(recall) against the False Positive rate in a plot across the different K values. Where the closer the Area under ROC curve(AUC) is to 1 the better is the classifier to distinguish between the different classes. Where if the AUC is at 0.5 the model's classifier is purely random.

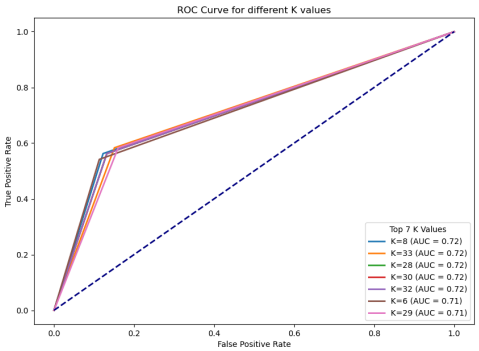

Figure 2: plot of Receiver Operating Characteristic

3 Conclusion

We concluded with that based on our predictions of our classes and the performance evaluation that followed, the most optimal fit for our model was the usage of the k value 8.