Chloe Schwieger (schwi415@umn.edu)
Maitrayee Deka (deka0031@umn.edu)

CSCI 5461 Homework 2

2. kNN performance evaluation

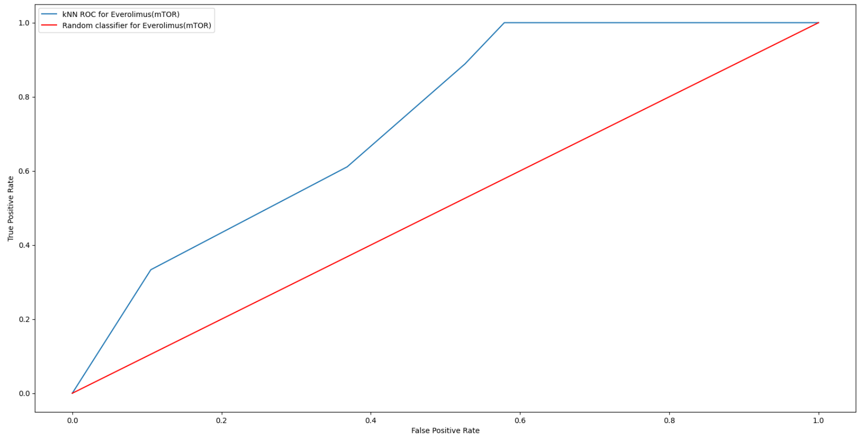
a. Does your classifier work better than a random classifier? For which drugs? Refer to specific evidence from your analysis to justify your answer.

For the drug 4-HC, the ROC curve appears to be random, as the curve is consistently below the random line which represents true positive rate (TPR) equals false positive rate (FPR). This suggests that the classifier's performance is even worse than guessing as there is still a higher FPR and lower TPR at any point where 0<x<1, with a smaller area under the curve (AUC).

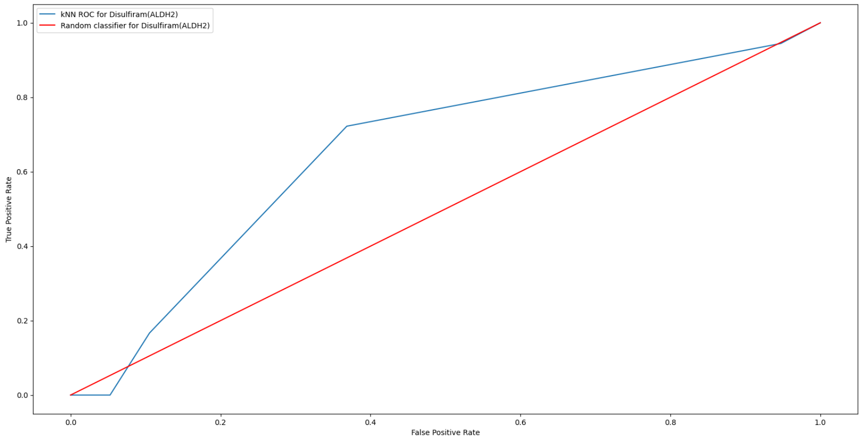
On the other hand, the ROC curves for Disulfiram, Everolimus, and Mebendazole are above the random line, indicating some level of accuracy. However, there is still a relatively high FPR, indicating that the classifier is not very reliable, but still better than random as there is more AUC.

Methylglyoxal's ROC curve seems to perform the best among the drugs, with a high TPR and low FPR at first, followed by a flat line as FPR catches up. This suggests that the classifier can more effectively identify actual positives as true positives, indicating a higher level of accuracy compared to the other drugs.

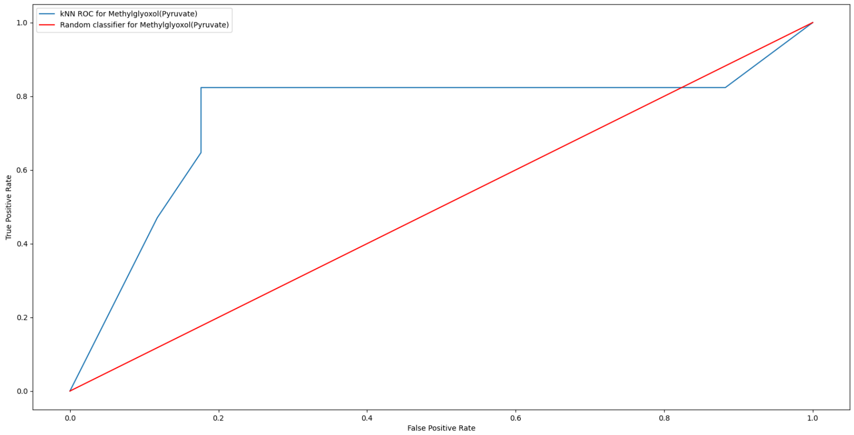
Everolimus



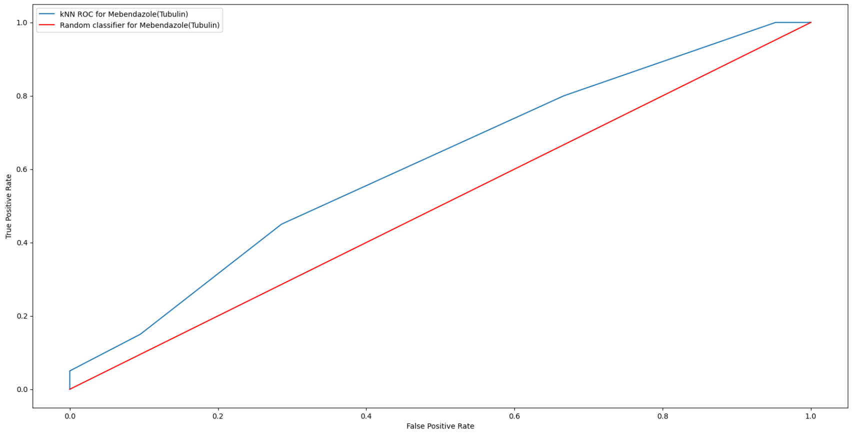
Disulfiram



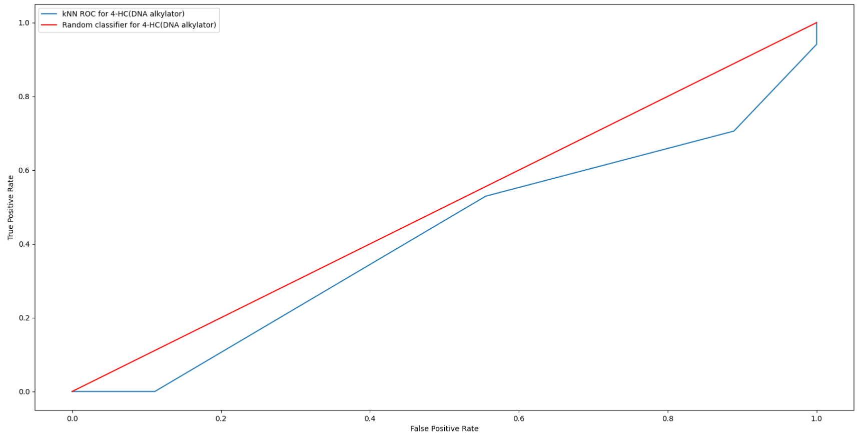
Methylglyoxal



Mebendazole



4-HC



b. For the drug on which your approach appears to work the best, how good is the classification performance? Pick a point on the ROC curve, and report the relevant metrics (number of true positives, false positives).

As stated before, Methylglyoxal's ROC curve seems to perform the best among the drugs, with a high TPR and low FPR. For example, the point (0.17647059, 0.82352941) corresponds to a particular threshold value for the classifier where the classifier correctly identifies over 82% of actual positive cases as positive and incorrectly identifies about 18% of actual negative cases as positive.

3. Exploration of parameters affecting kNN performance

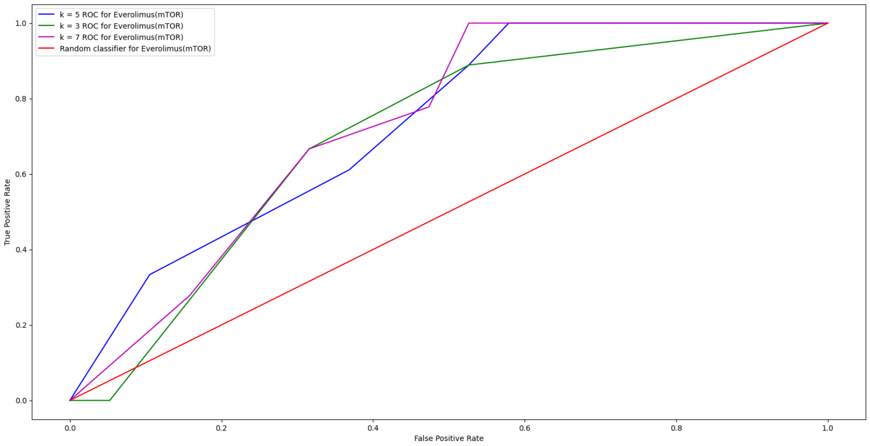
a. Discuss the results of your analysis. Does the choice of k affect the performance of the classifier?

For this particular dataset, the choice of k value in the KNN classifier did not seem to have a significant impact on the performance of the classifier. Whether k was set to 3, 5, or 7, the ROC curves were relatively similar, indicating that the performance of the classifier seemed consistent across different k values.

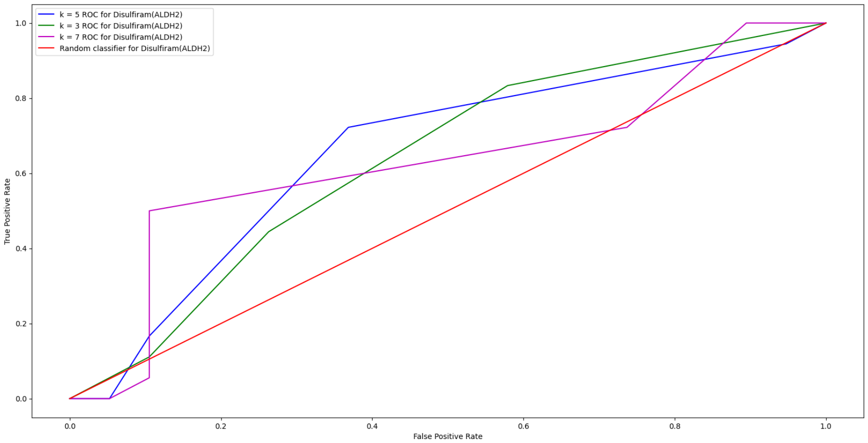
However, it is worth noting that the ROC curve for k=7 had a noticeably less smooth curve and dipped below the random line at one point with Disulfiram and Mebendazole. Also, k=7 was notably more random than k=3 or k=5 in 4-HC the majority of the time. These behaviors can be explained as when the k value is too large, the model may become overly simplified and unable to capture the nuances of the underlying gene expression data. This can lead to underfitting, where the model is too simple to accurately capture the relationships between the gene expression patterns and drug sensitivity classification.

k=3 k=5 k=7

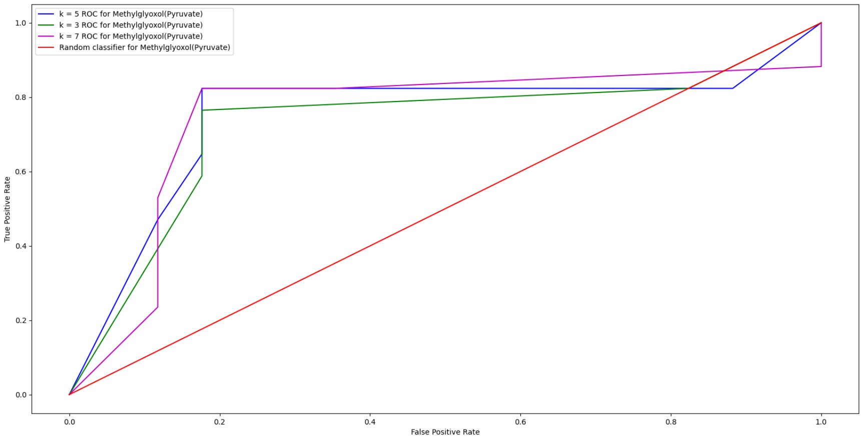
Everolimus



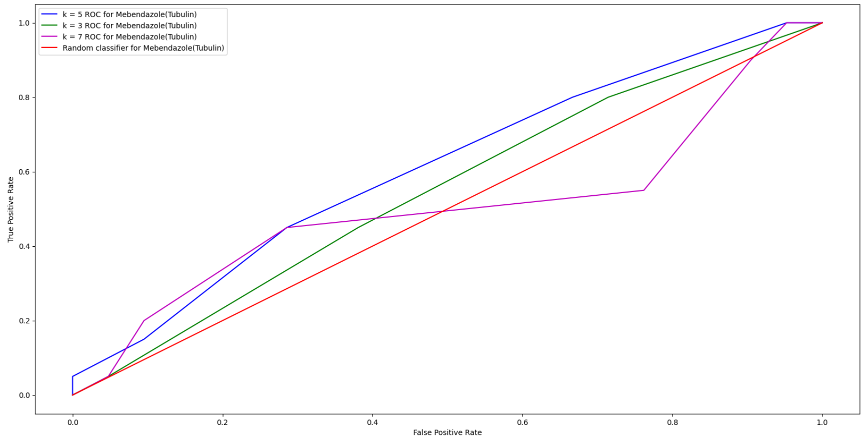
Disulfiram



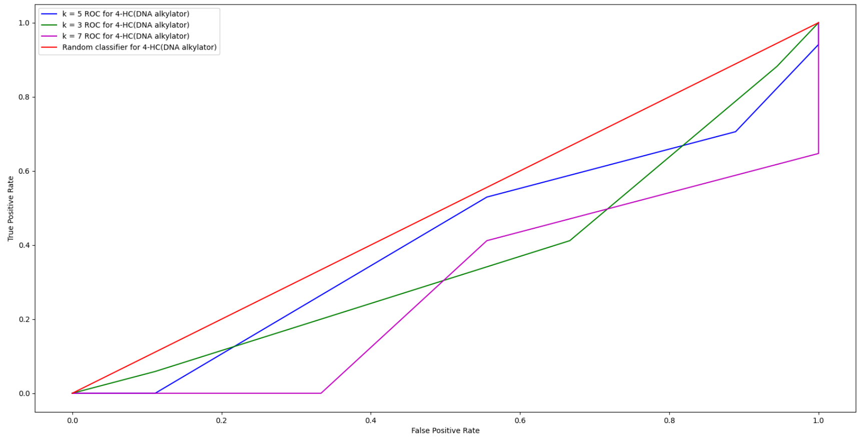
Methylglyoxal



Mebendazole

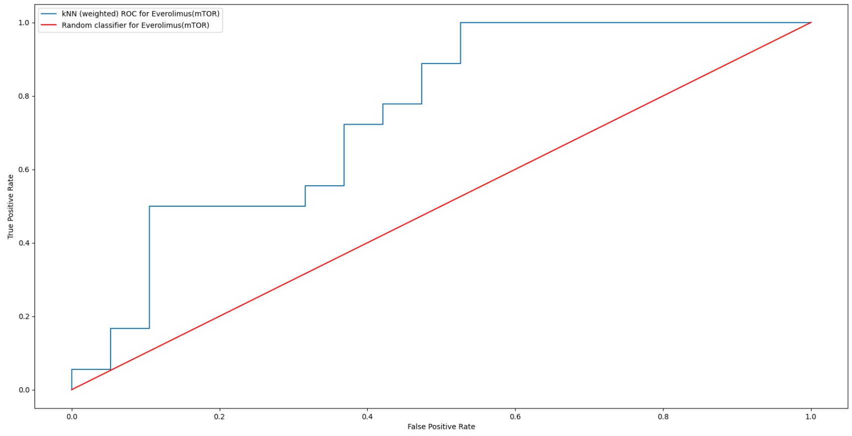


4-HC

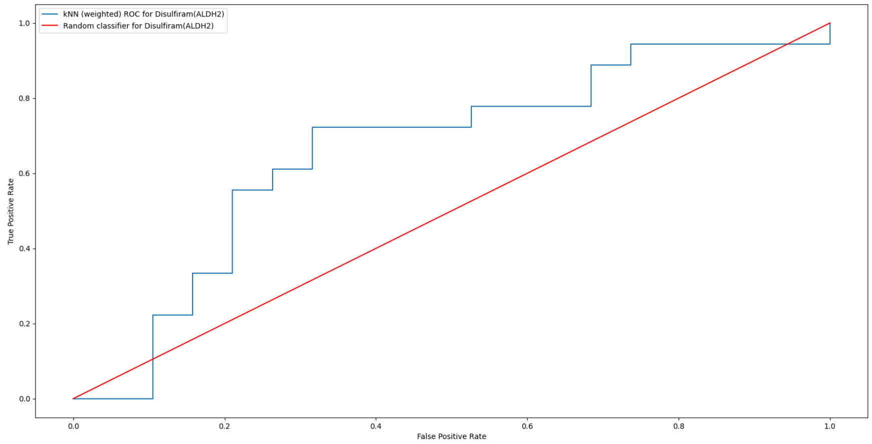


b. Weighted Plots

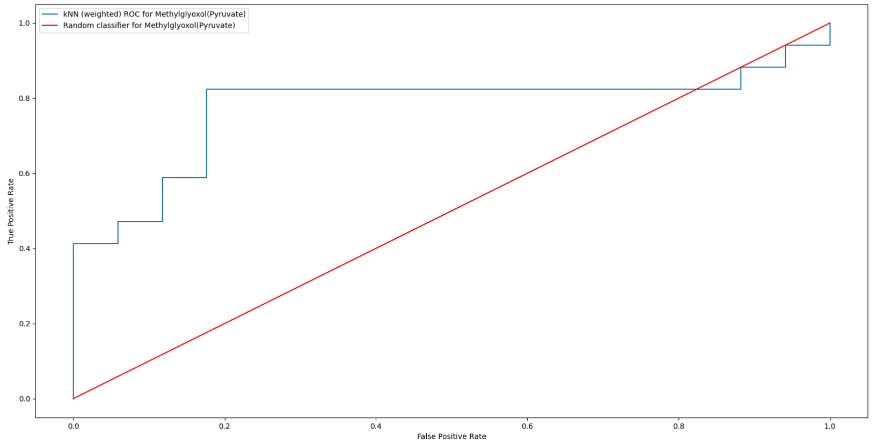
Everolimus



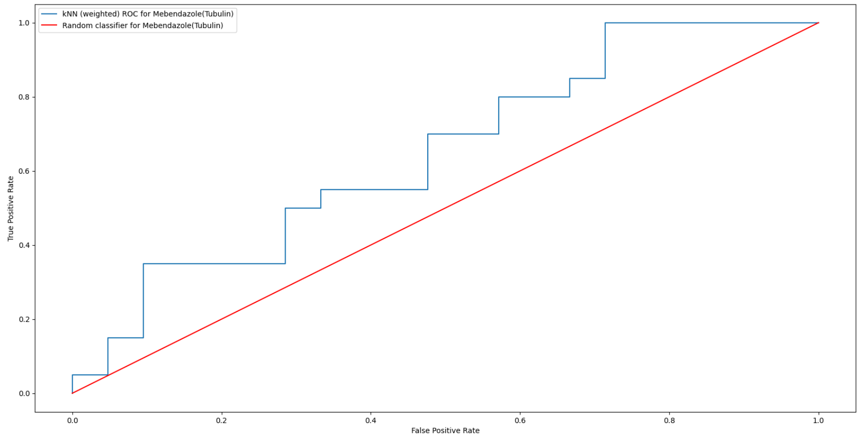
Disulfiram



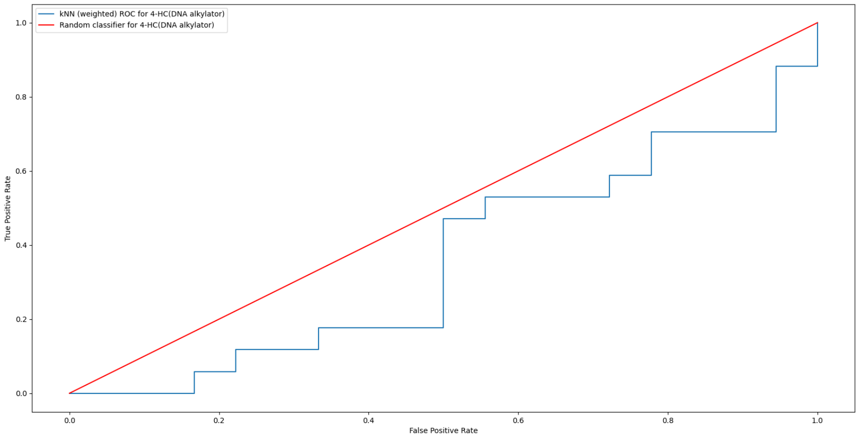
Methylglyoxal



Mebendazole



4-HC



As can be seen from the ROC plots above, for each drug, using the weighted score leads to more step-wise variation in the ROC. However, the same overall patterns (compared to the ROCs from Problem 2) can be observed. In general, the ROCs in this scenario tend to dip above and below the random classifier performance in the same areas as before.

For the drug 4-HC, we continue to see the worst performing classifier, since the classifier appears to consistently perform worse than a random classifier across all values of TPR and FPR.