Turgut Alp Edis

21702587

**GE461 – Introduction to Data Science**

**Project 5**

**Comparing Different Classification Algorithms on Hyperplane Dataset**

**1 – Dataset Generation**

In this project, the dataset is created using the Hyperplane Generator provided by the sckit-multiflow.data module. Four different datasets are created and each of them has 20000 instances and each instance has 10 features. As it indicated in the project description, first dataset is created with 0.1 noise and 2 drifting features, second is created with 0.3 noise and 2 drifting features, third dataset is created with 0.1 noise and 5 drifting features and last dataset is created with 0.3 noise and 5 drifting features. Then, they are named according to their noise percentage and number of drifting features, for instance “Hyperplane Dataset 10\_2.csv”.

**2 - Data Stream Classification with Three Separate Online Single Classifiers: HT, KNN, NB**

Three classifiers are used in this project, which are Hoeffding Tree (HT), K-Nearest Neighbor (KNN) and Naive Bayes (NB). These three classifiers do not require any other parameters to work properly, each of them can work properly in default. In the figure 1, 2, 3 and 4, it can be seen three classifiers on four different datasets.

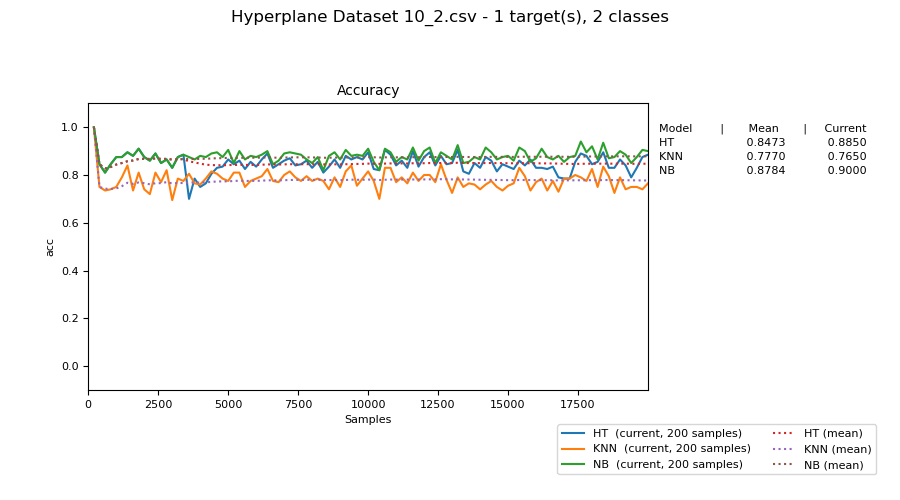


Figure 1. Hyperplane Dataset 10\_2 Accuracy for Three Classifier

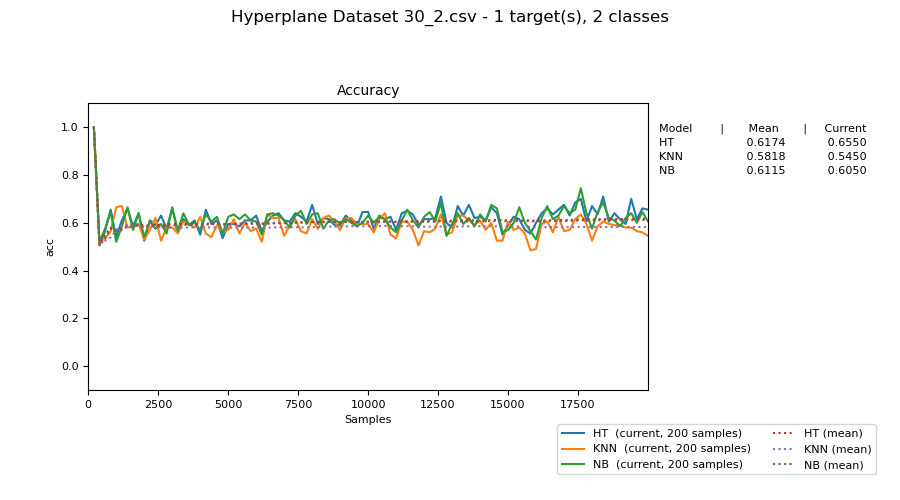


Figure 2. Hyperplane Dataset 30\_2 Accuracy for Three Classifier

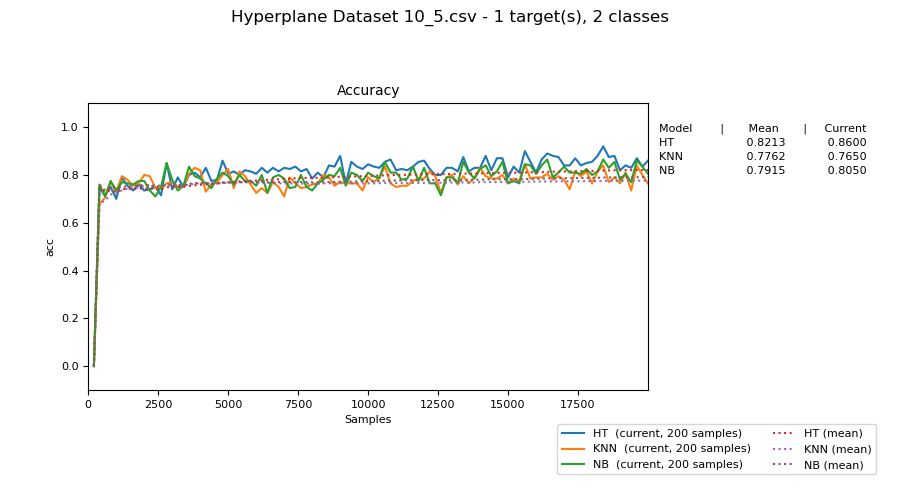


Figure 3. Hyperplane Dataset 10\_5 Accuracy for Three Classifier

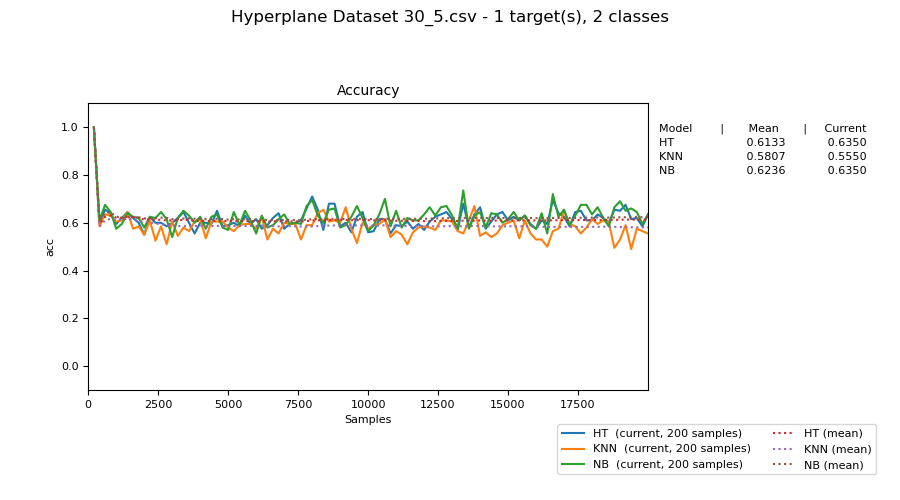


Figure 4. Hyperplane Dataset 30\_5 Accuracy for Three Classifier

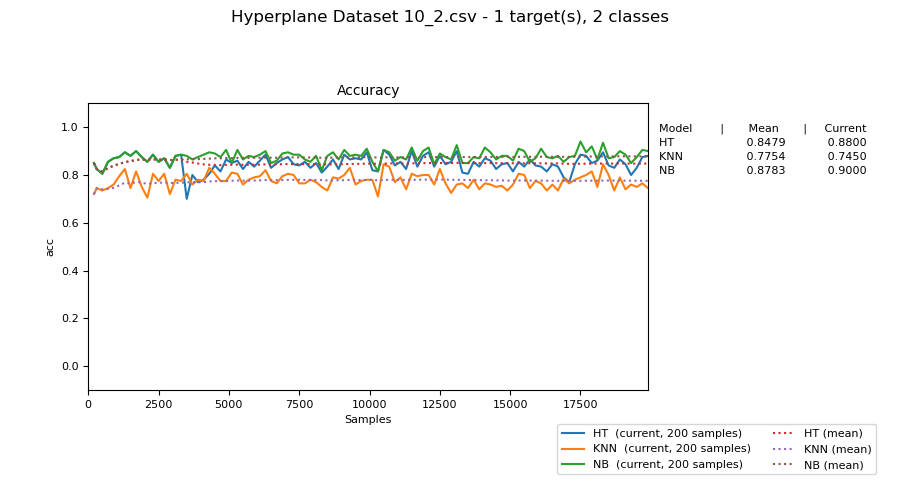


Figure 5. Hyperplane Dataset 10\_2 Accuracy for Three Classifier Batch Size = 10

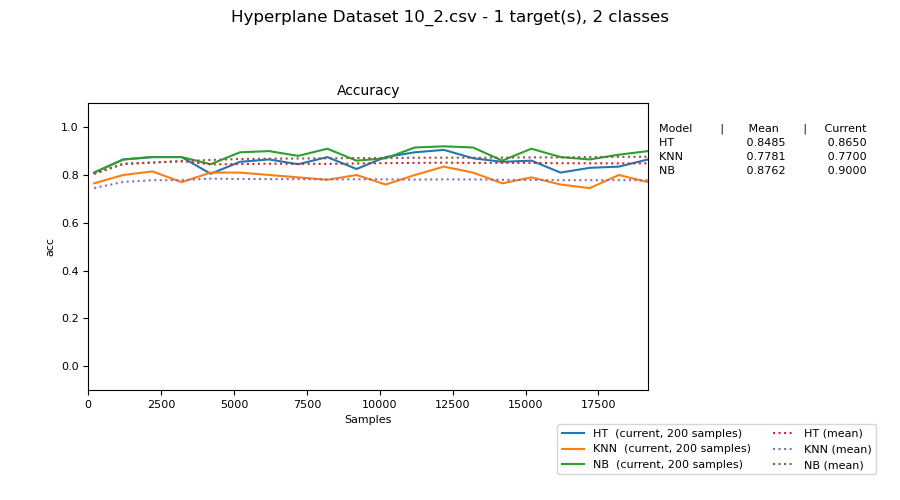


Figure 6. Hyperplane Dataset 10\_2 Accuracy for Three Classifier Batch Size = 100

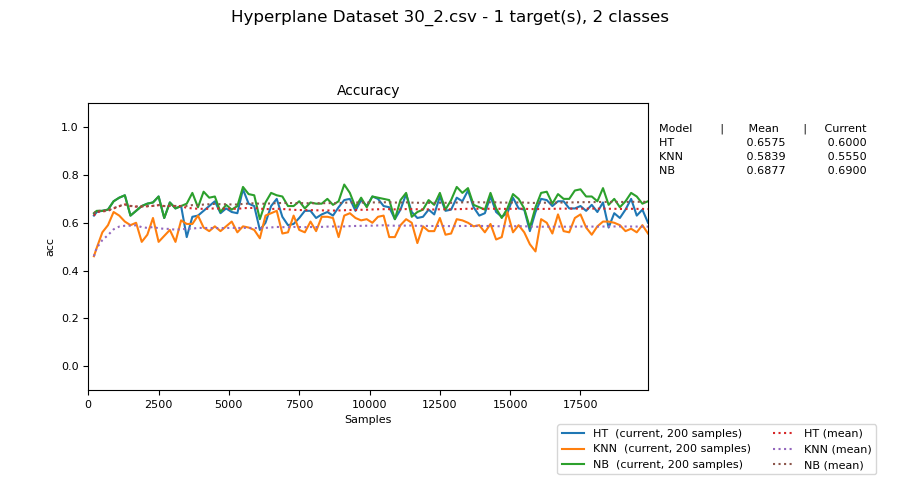


Figure 7. Hyperplane Dataset 30\_2 Accuracy for Three Classifier Batch Size = 10

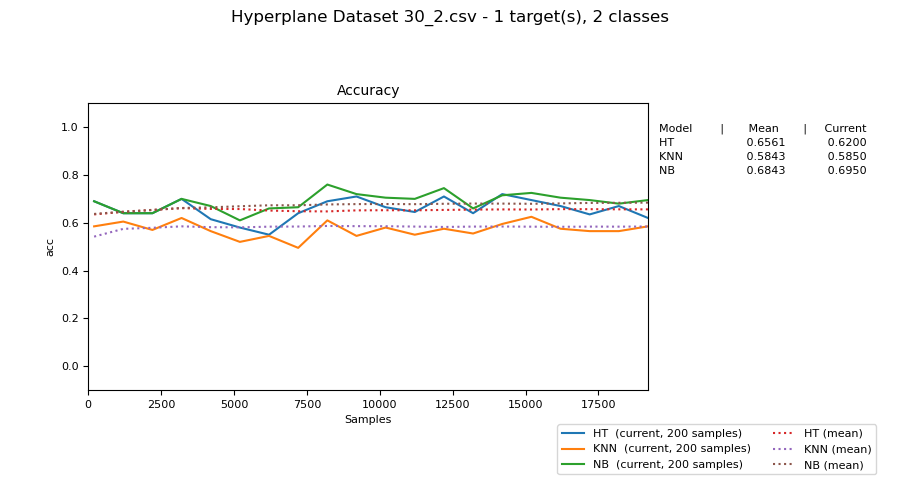


Figure 8. Hyperplane Dataset 30\_2 Accuracy for Three Classifier Batch Size = 100

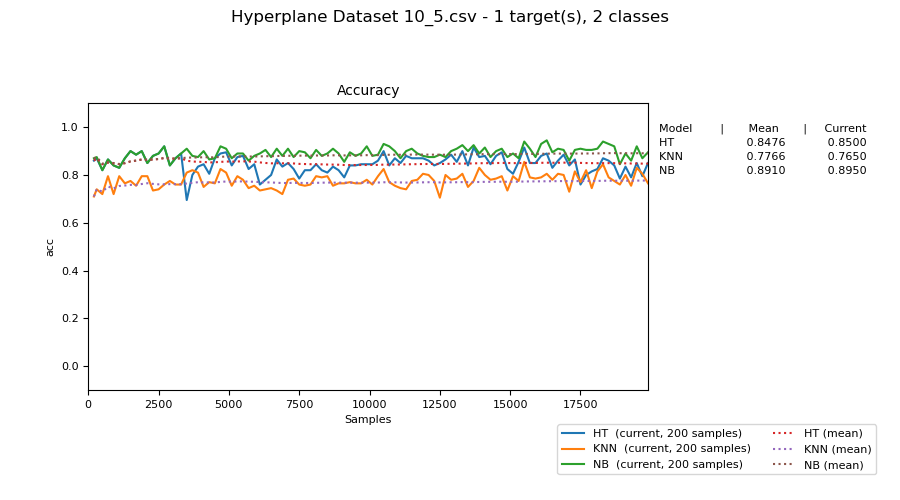


Figure 9. Hyperplane Dataset 10\_5 Accuracy for Three Classifier Batch Size = 10

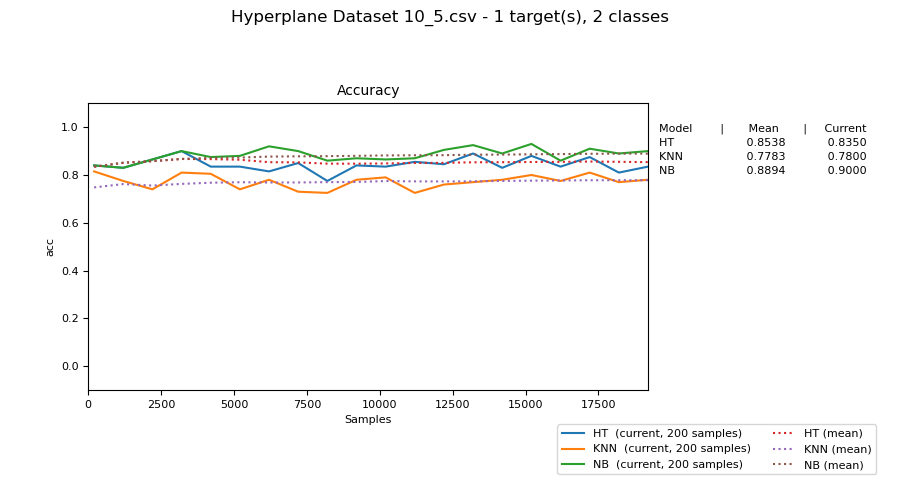


Figure 10. Hyperplane Dataset 10\_5 Accuracy for Three Classifier Batch Size = 100

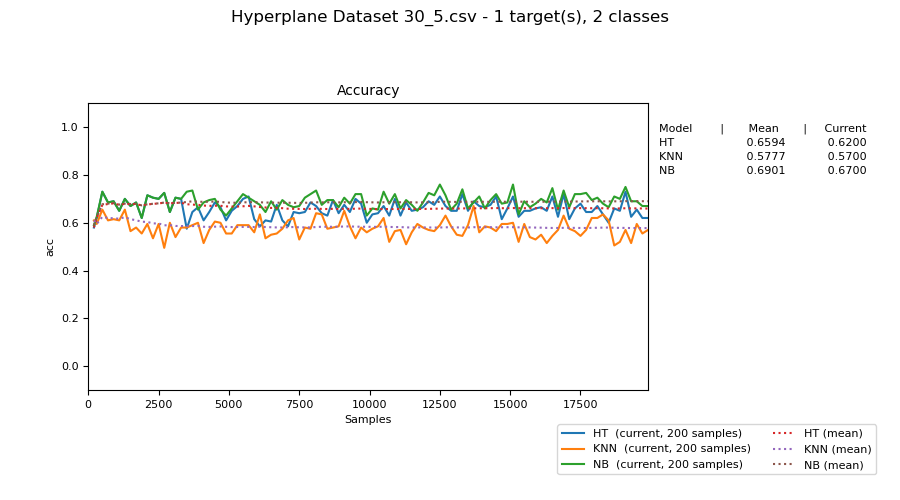


Figure 11. Hyperplane Dataset 30\_5 Accuracy for Three Classifier Batch Size = 10

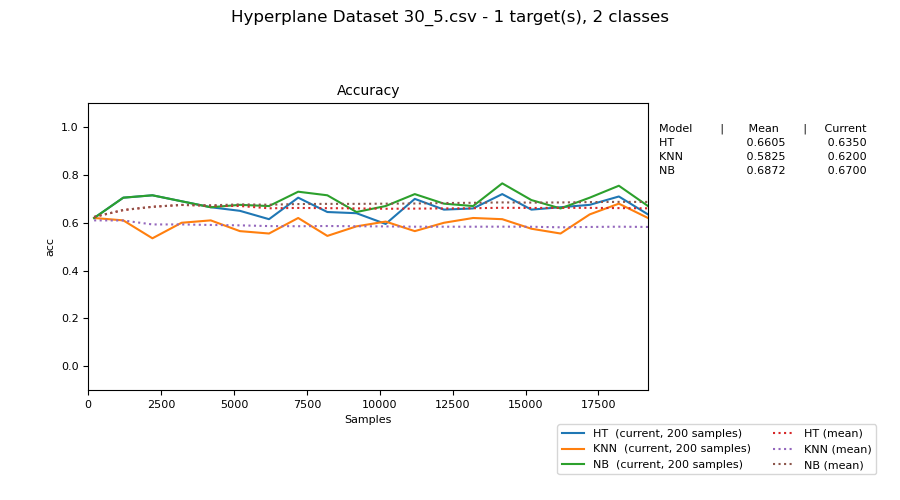


Figure 12. Hyperplane Dataset 30\_5 Accuracy for Three Classifier Batch Size = 100

From the figure 1 and 3, it can be said that when the drifting features increases, the accuracy of HT and NB decreases. The reason could be that when the dataset is created, some values drifted to the beginning so that they lower the accuracy. However, even if their accuracies decreases, they are still above the 75 percent, so it can be stated that they are high.

On the other hand, from the figure 2 and 4, it can be said that when the drifting features increases, only HT decreases. However, their accuracy values are almost the same, so this difference is not important.

From the other figures, it can be inferred that the batch size makes the accuracy more stable, therefore it is better when their values increase.