CSI5155 Machine Learning
Assignment 4

Introduction

In this assignment, we are asked to use three Insects data streams (Abrupt, Incremental and Gradual) and perform online learning using scikit multiflow. We perform experiments using the following classifiers and compare and contrast their results with the reference paper:

- No Change Classifier
- Majority Class Classifier
- Hoeffding Trees Classifier
- SAM-KNN Classifier
- Hoeffding Adaptive Trees Classifier
- Accuracy Weighted Ensemble Classifier
- Adaptive Random Forest Classifier
- Hoeffding Tree Classifier with Drift Detection

No Change Classifier and Majority Class Classifier

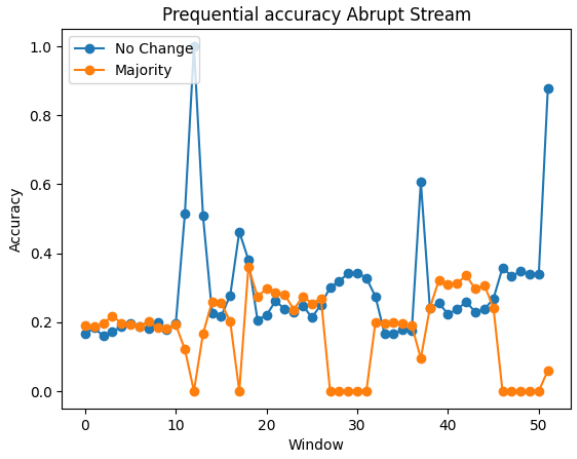
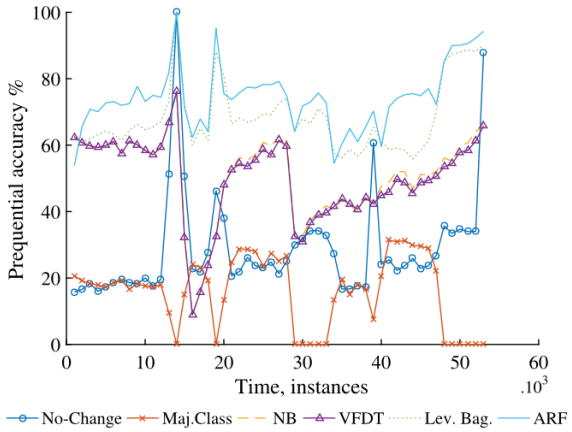
We use No Change Classifier and Majority Class classifier with a sliding window of 1000 to calculate the prequential accuracy and plot the prequential accuracy similar to the reference paper. The results obtained are as follows:

	Experiment Results		**Reference Paper Results**	
	No Change	**Majority**	**No Change**	**Majority**
Abrupt	29.23	17.60	28.98	16.07
Gradual	38.23	15.95	38.43	15.76
Incremental	16.05	16.96	16.04	11.51

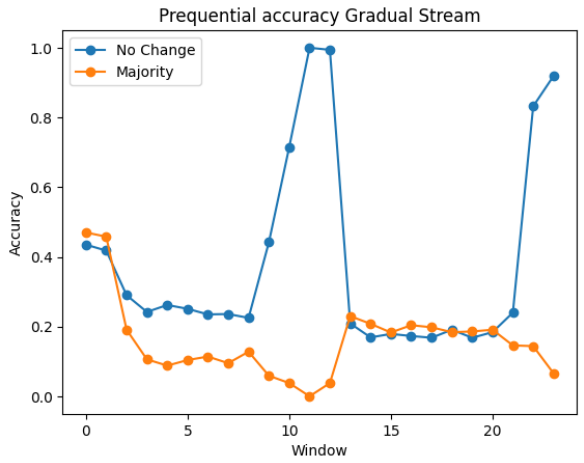
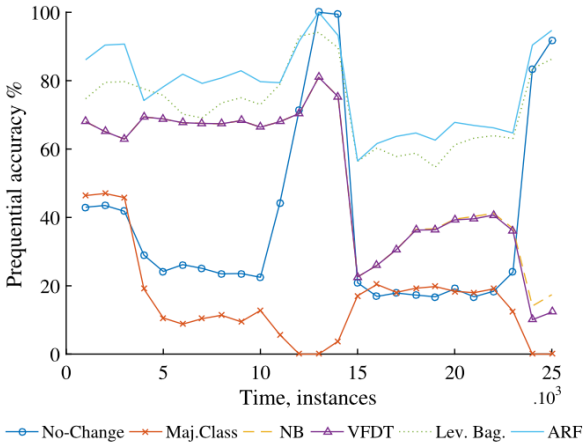
The comparison plots between the experiment results and Reference paper are as follows:

Surya Kiran Suresh (300269939)

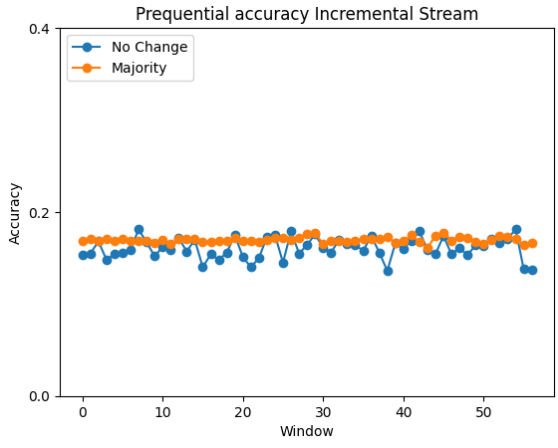
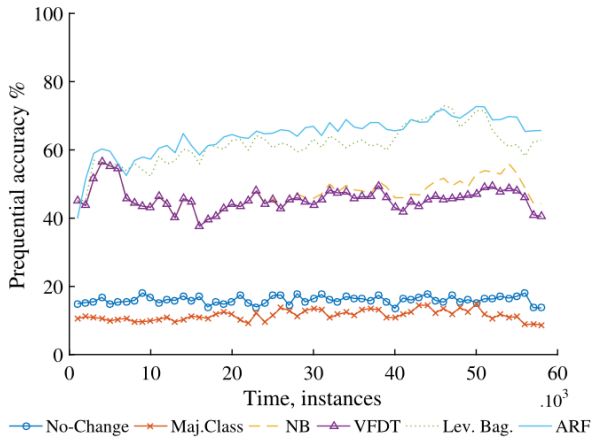
Prequential Accuracy for Abrupt Stream Data

Experimentation	Reference Paper
	

Prequential Accuracy for Gradual Stream Data

Experimentation	Reference Paper
	

Prequential Accuracy for Incremental Stream Data

Experimentation	Reference Paper
	

Surya Kiran Suresh (300269939)

If we compare the results, we can find that both the results of the reference paper and our experimentation are comparable and the plots look similar for no change and majority class classifier. We find **a difference in accuracy** only in **majority class classifier for Incremental Stream Data**.

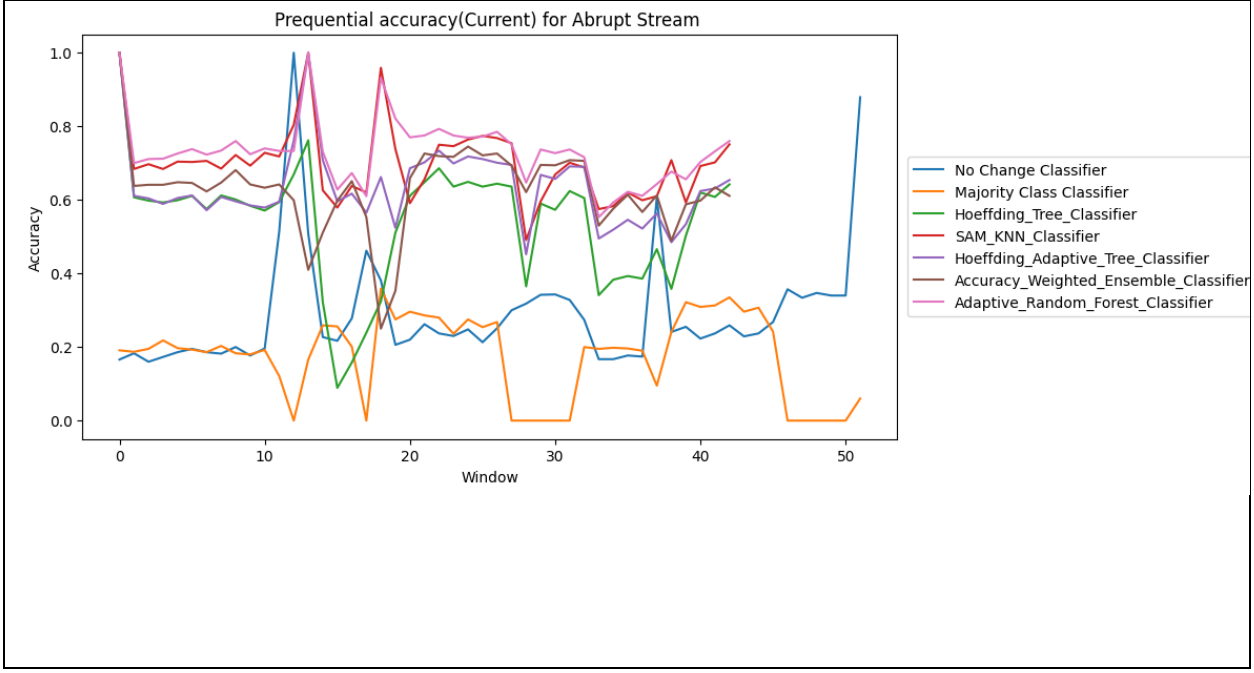
Hoeffding Tree, SAM-KNN, Hoeffding Adaptive Tree and 2 Ensemble based Classifiers

In this section, we will be looking into the prequential accuracies of 5 different models on the 3 streams of data we have. The following are the results obtained:

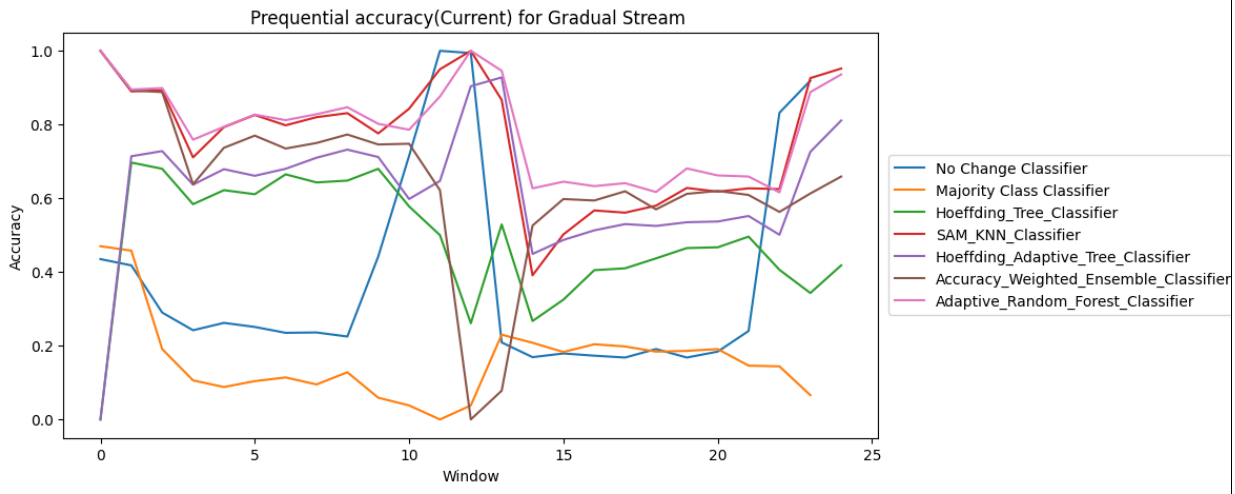
	Hoeffding Tree	SAM KNN	Hoeffding Adaptive Tree	Accuracy Weighted Ensemble	Adaptive Random Forest
Abrupt	52.91	69.21	62.77	61.80	72.46
Gradual	51.01	74.18	64.02	62.21	77.28
Incremental	54.12	60.69	54.78	59.64	64.73

The algorithm comparison plots for the different data streams are as follows:

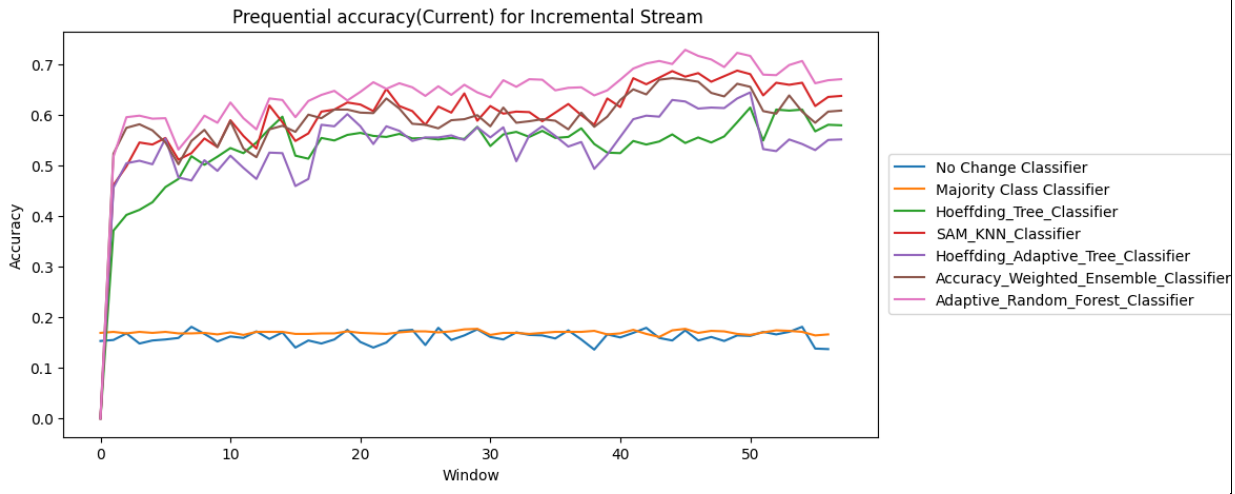
Abrupt Stream



Gradual Stream



Incremental Stream



It is observed that we get better results than the no change classifier and majority class classifier for the 3 data streams when we use aforementioned algorithms.

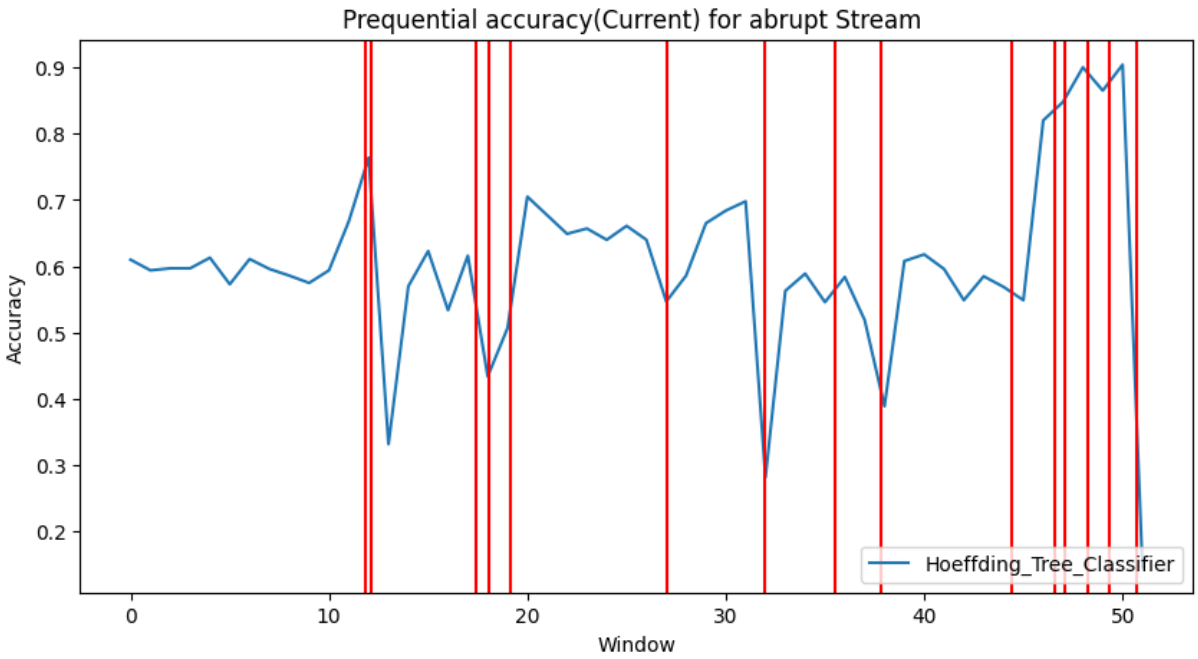
Hoeffding Tree Classifier with Drift Detection

Since the data stream that we are working on is susceptible to **data and concept drift**, if we employ a **drift detection method** along with the classifier, we can achieve better results. For this part of the experimentation, we use **ADWIN concept drift detection** method to find out changes in data and if there are any, we **trigger a retraining pipeline** to improve the prequential accuracy of the model. The following results table is a comparison between Hoeffding Tree Classifier with and without drift detection:

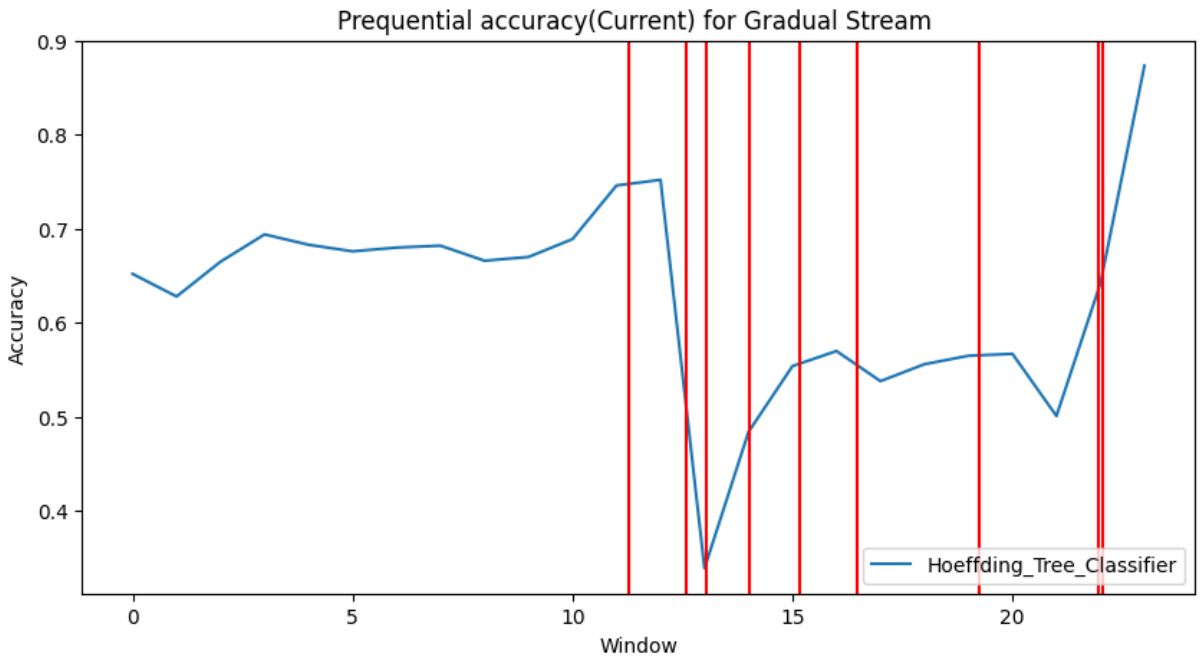
	Hoeffding Tree	Hoeffding Tree with ADWIN Drift Detection	Number of windows with drift
Abrupt	52.91	60.44	15
Gradual	51.01	62.80	9
Incremental	54.12	55.91	12

If we notice the table above, we can find that there is a significant improvement in the Prequential Accuracy from naïve Hoeffding Tree classifier when compared to the one coupled with the drift detection model. Once we detect a drift, we **drop the previous model** and **retrain a new model based on the last 1000 records** in the data.

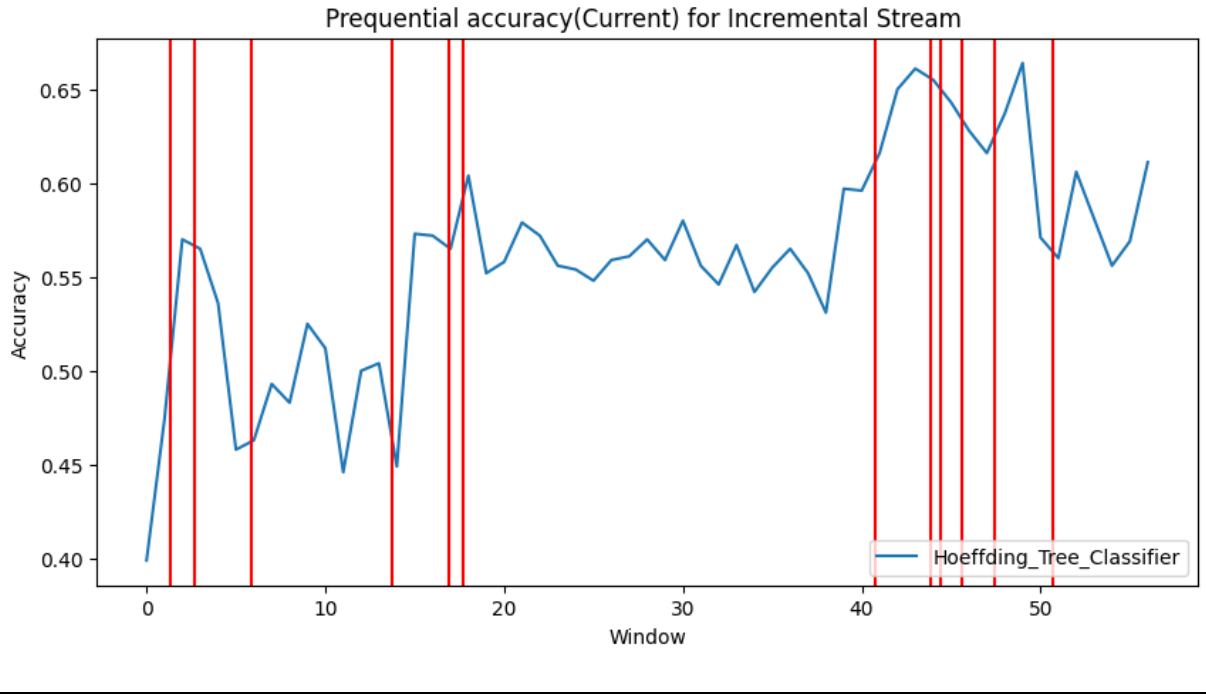
The table below shows the prequential accuracy along with the drift detections for the different streams of data:

Abrupt Stream


Gradual Stream



Incremental Stream



Summary Table

The following table shows the prequential accuracies for all the models that were experimented in this assignment:

	No Change Classifier	Majority Class Classifier	Hoeffding Tree	SAM KNN	Hoeffding Adaptive Tree	Accuracy Weighted Ensemble	Adaptive Random Forest	Hoeffding Tree with Drift Detection
Abrupt	29.23	17.60	52.91	69.21	62.77	61.80	72.46	60.44
Gradual	38.23	15.95	51.01	74.18	64.02	62.21	77.28	62.80
Incremental	16.05	16.96	54.12	60.69	54.78	59.64	64.73	55.91

Results and Lessons learnt

In this assignment we were able to learn how to handle data streams and to use them for online learning. We were able to **reproduce the results** from the **reference paper for no change and majority class classifier**. The other models that were used in this experiment were different from the paper. We also **learnt how to detect changes** in the data using statistical methods and perform actions based on those changes to improve the accuracy of the model.

Contrast with the Reference Paper

- As mentioned in the previous sections we were able to reproduce the results of no change and majority class classifier
- The entire experiment was conducted in the same setting as the paper (ie) with a **sliding window of 1000** with a **pretrain window of 1000**
- For Drift detection model, we used **Hoeffding Tree model** instead of the **Naïve Bayes model** used in the **reference paper**. The drift window in the reference paper was set as 100 but, in this experimentation, we **feed the data one by one** for each prequential accuracy window of 1000 records.
- In every window, if we detect a change, we mark that point and don't explore the window further for any changes and trigger the retraining pipeline.