A REPORT
On

Assignment 2 – Active Learning

By

Varkeychan Jacob	2017B5A70828P
Aswin Benedict	2019A4PS0579P

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
(DECEMBER 2021)

Table of Contents

Introduction

The objective of the classifier is to classify emails into 6 distinct labels of:

- Academics
- Miscellaneous
- News
- Placements
- Promotions
- Social

The dataset is the same as the one used in the previous set, procured from our personal mail accounts. They are stored in the form of a CSV file.

The assignment is programmed in Python. All the classifiers were imported from the scikit-learn library.

The dataset is initially partitioned into training and testing sets and the training set is further partitioned into 10% initial labelled data and 90% active training data.

For Uncertainty Sampling, all four methods of least confidence, smallest margin, largest margin and entropy were tried out. The different disagreement strategies were implemented with help of modAL framework.

For Query-by-Committee both Vote Entropy and KL-Divergence were used.

Query-by-Committee was implemented with bagging on 5 different classifiers.

The different classifiers used were:

- Decision Tree
- Random Forest
- Linear SVM
- RBF SVM
- Sigmoid SVM

All the active learning models were tried out using 6 different classifiers and graphs were plotted including the passive learning performance for the corresponding amount of labelled data. The graphs have x-axis as percentage of actively learnt data and y-axis as the corresponding accuracy.

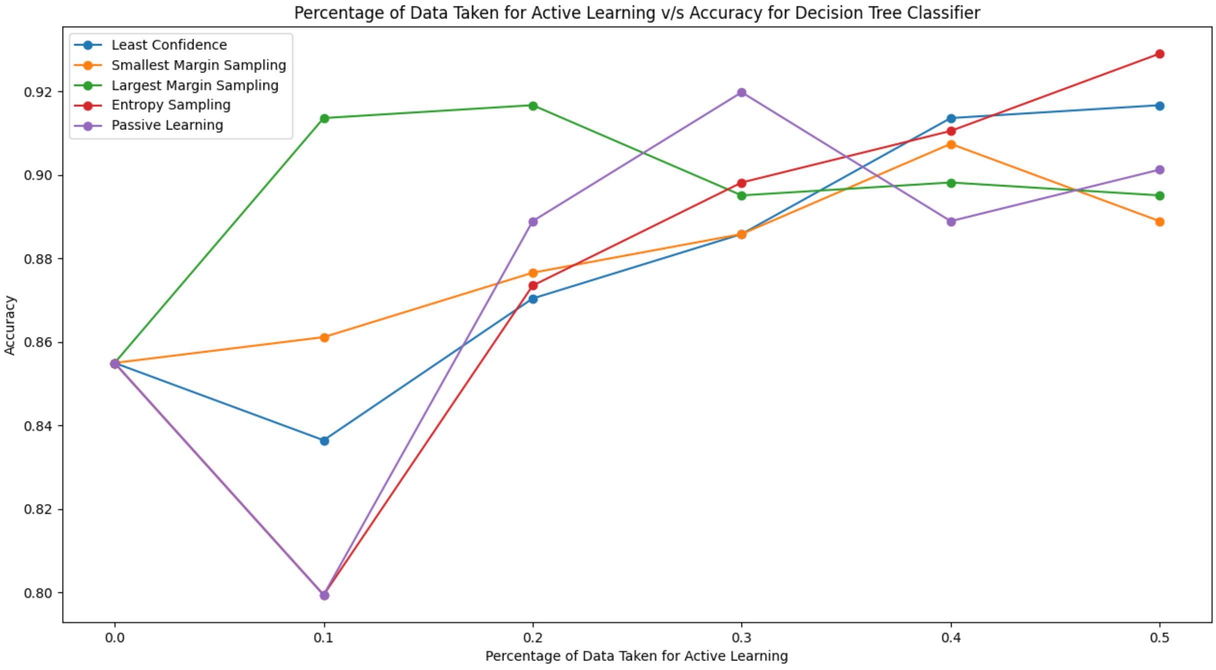
The different classifiers used on the Active Learning models:

- Decision Tree
- Random Forest
- Linear SVM
- RBF SVM
- Sigmoid SVM
- Polynomial SVM

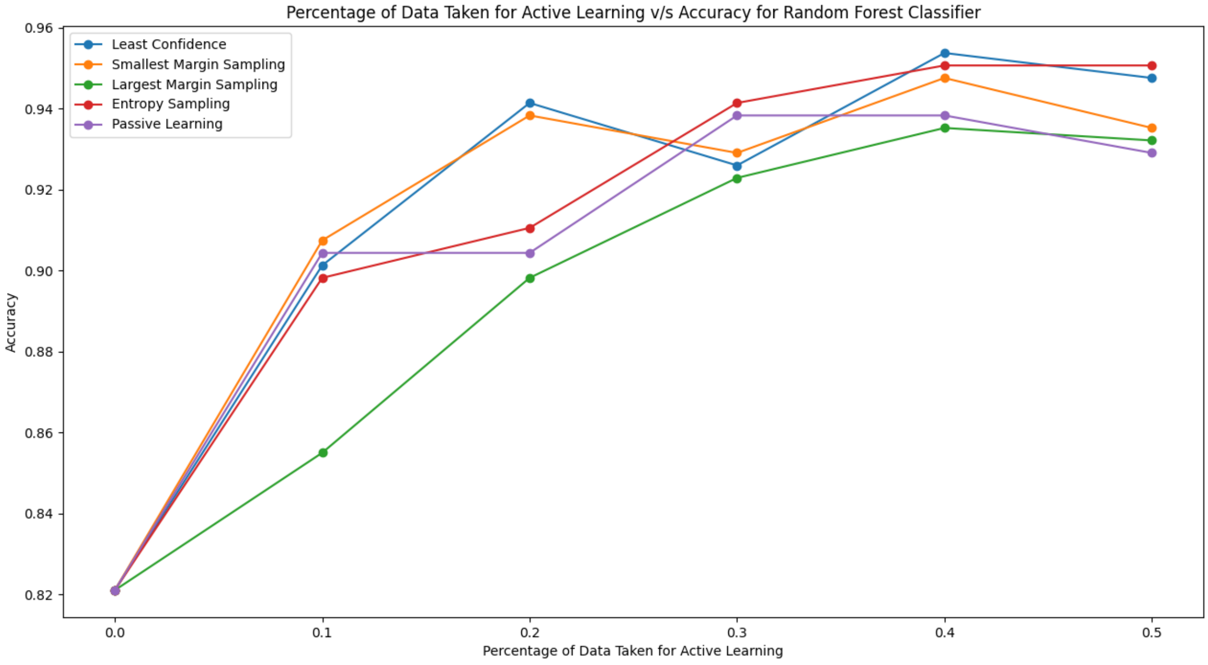
Results

1. Uncertainty Sampling

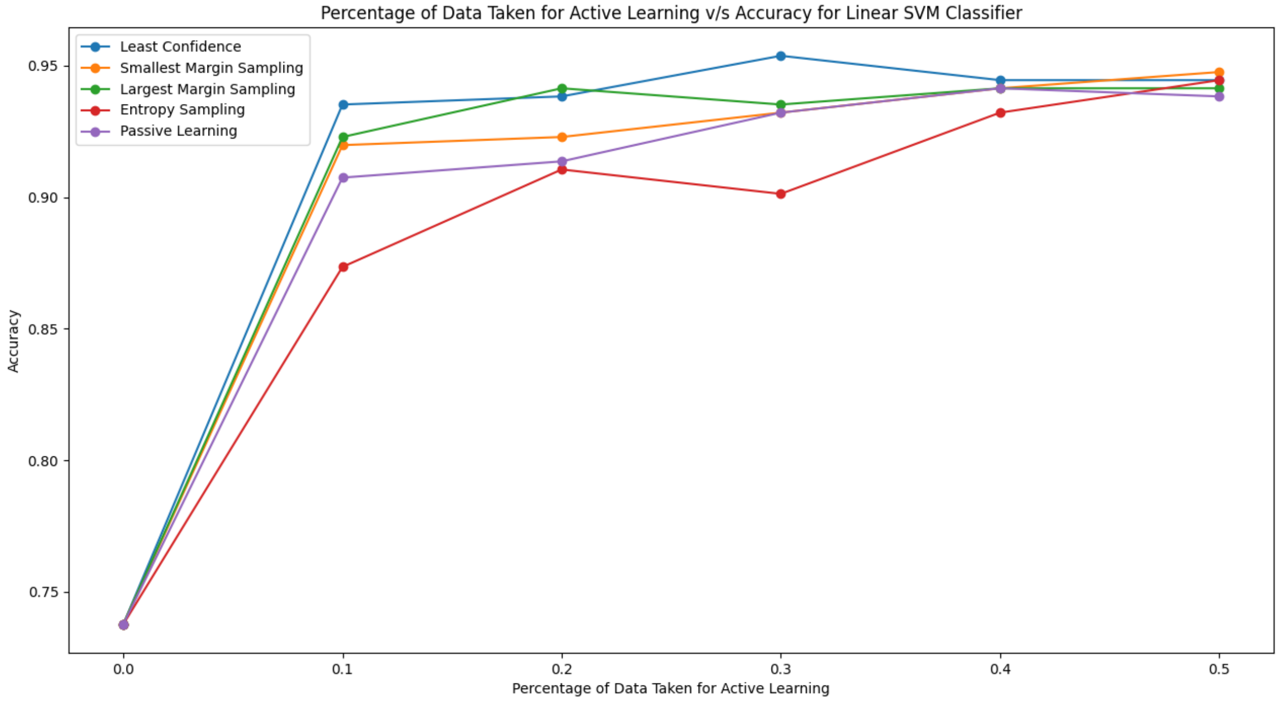
1.1 Decision Tree Classifier



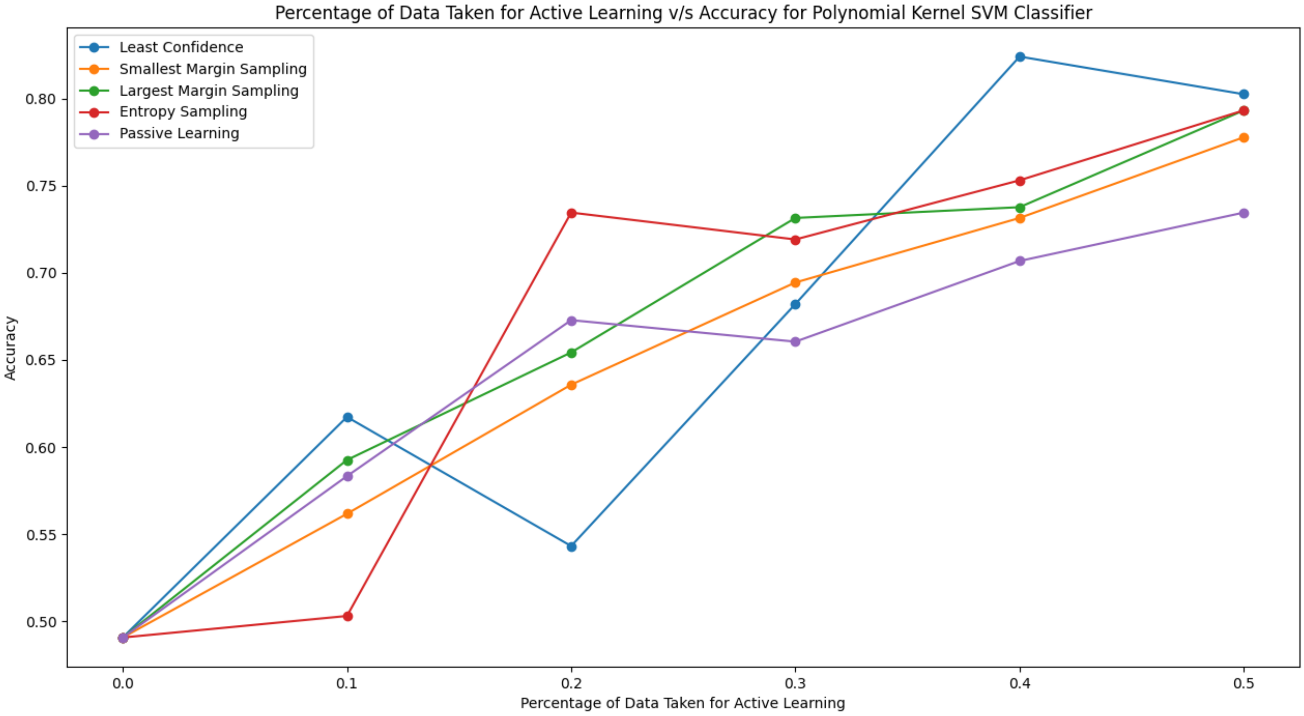
1.2 Random Forest Classifier



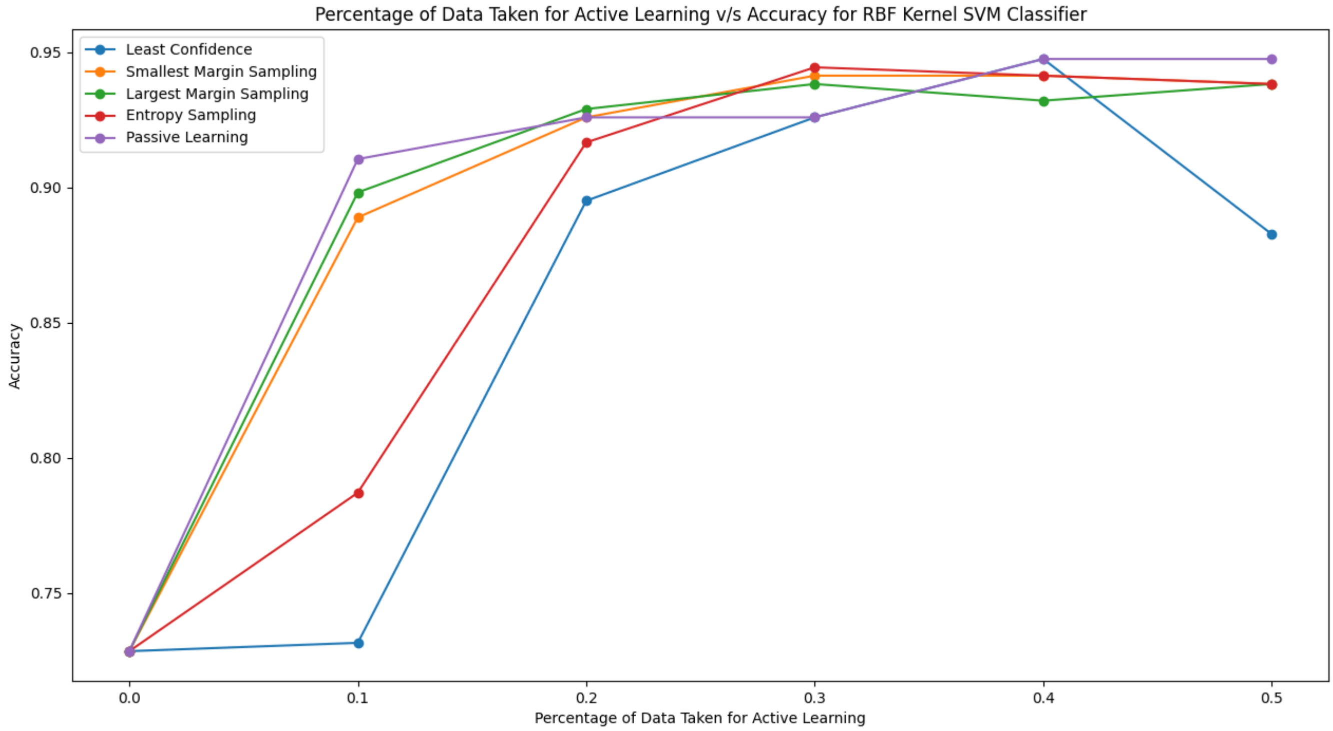
1.3 Linear SVM



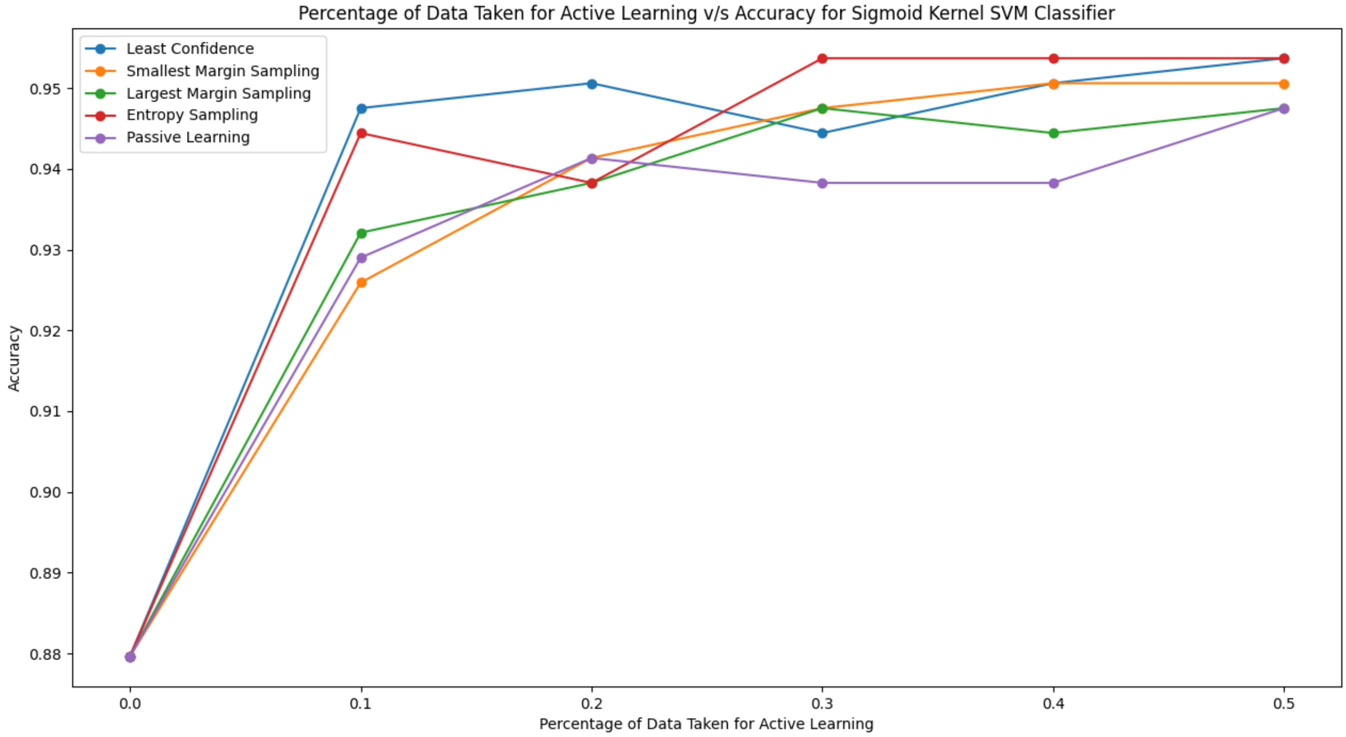
1.4 Polynomial SVM



1.5 RBF Kernel SVM

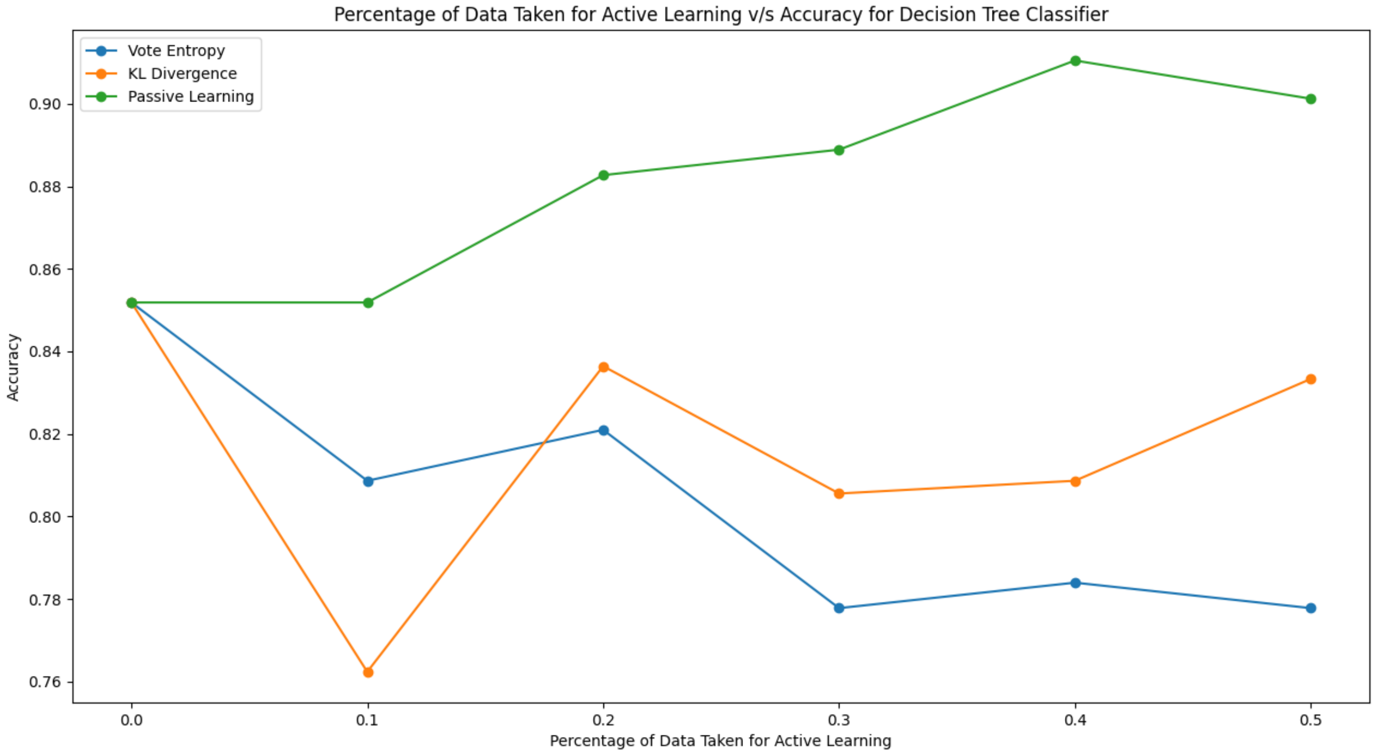


1.6 Sigmoid SVM

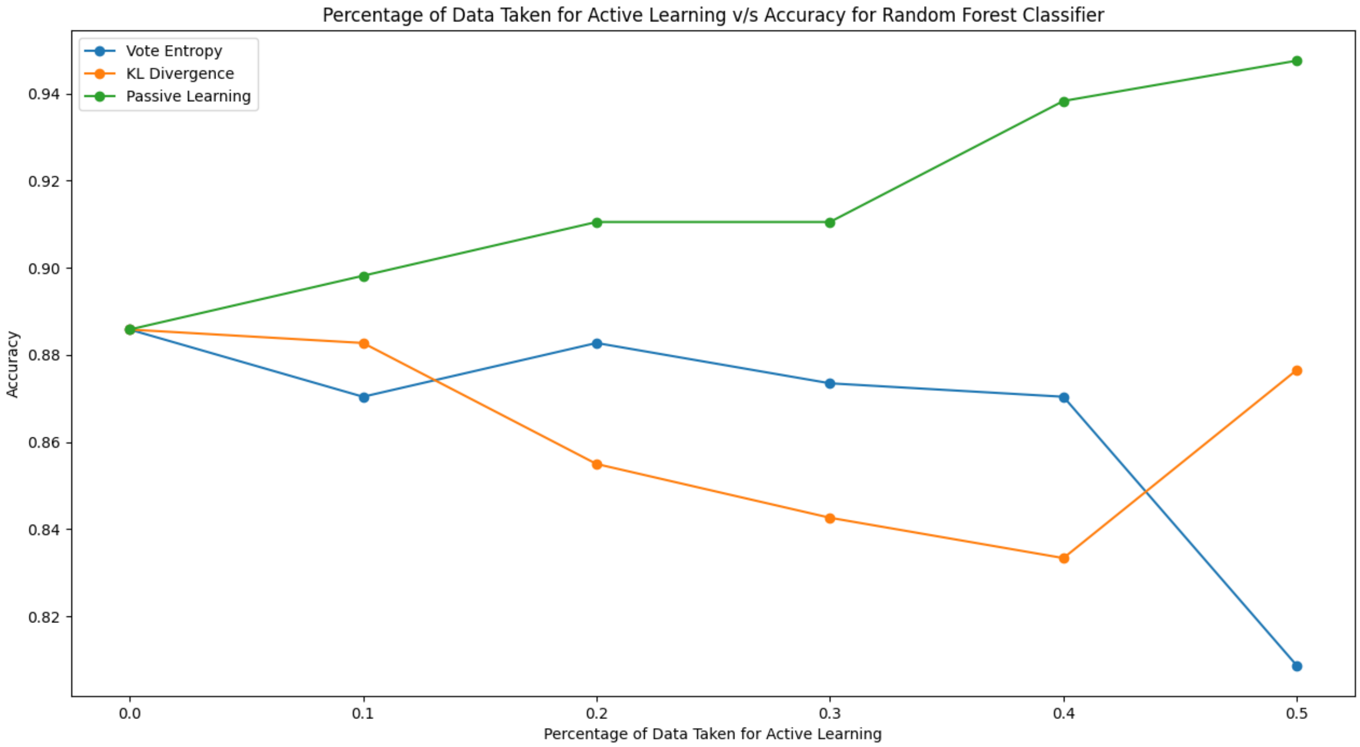


2. Query-by-Committee

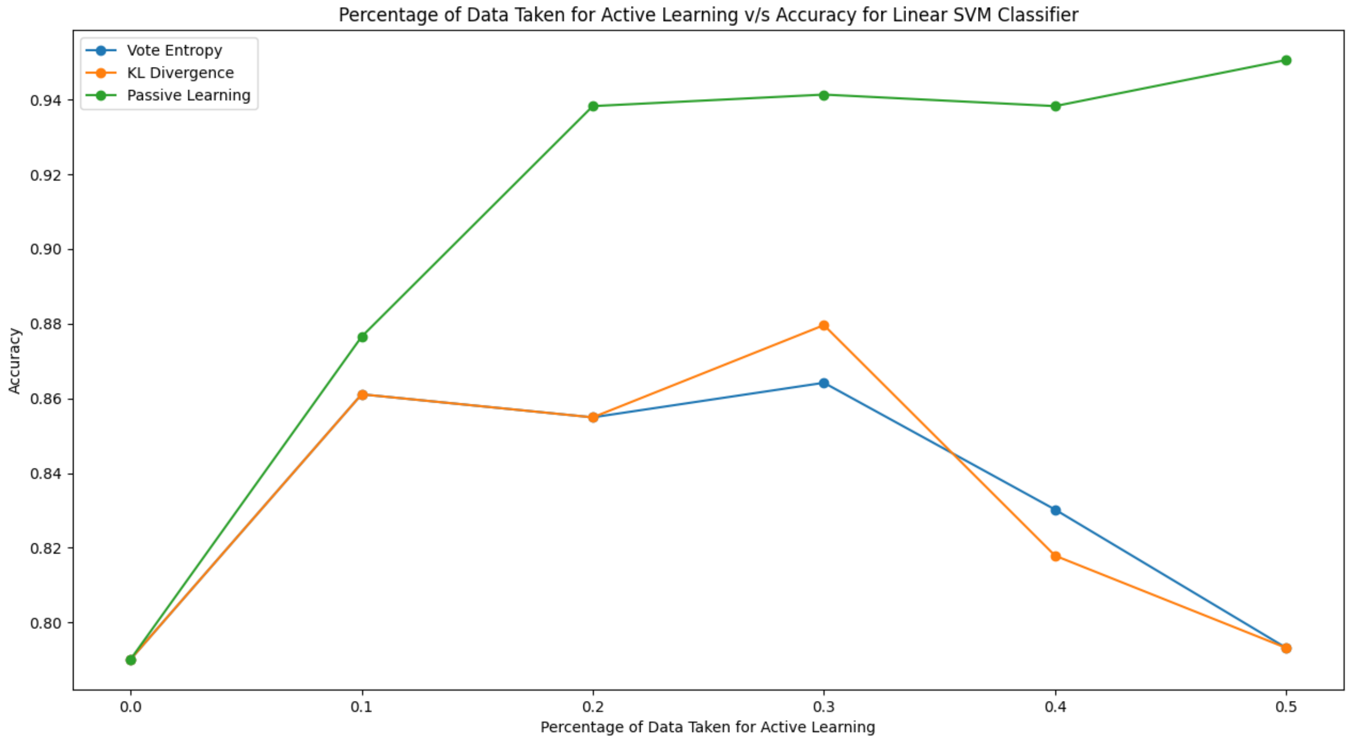
2.1 Decision Tree Classifier



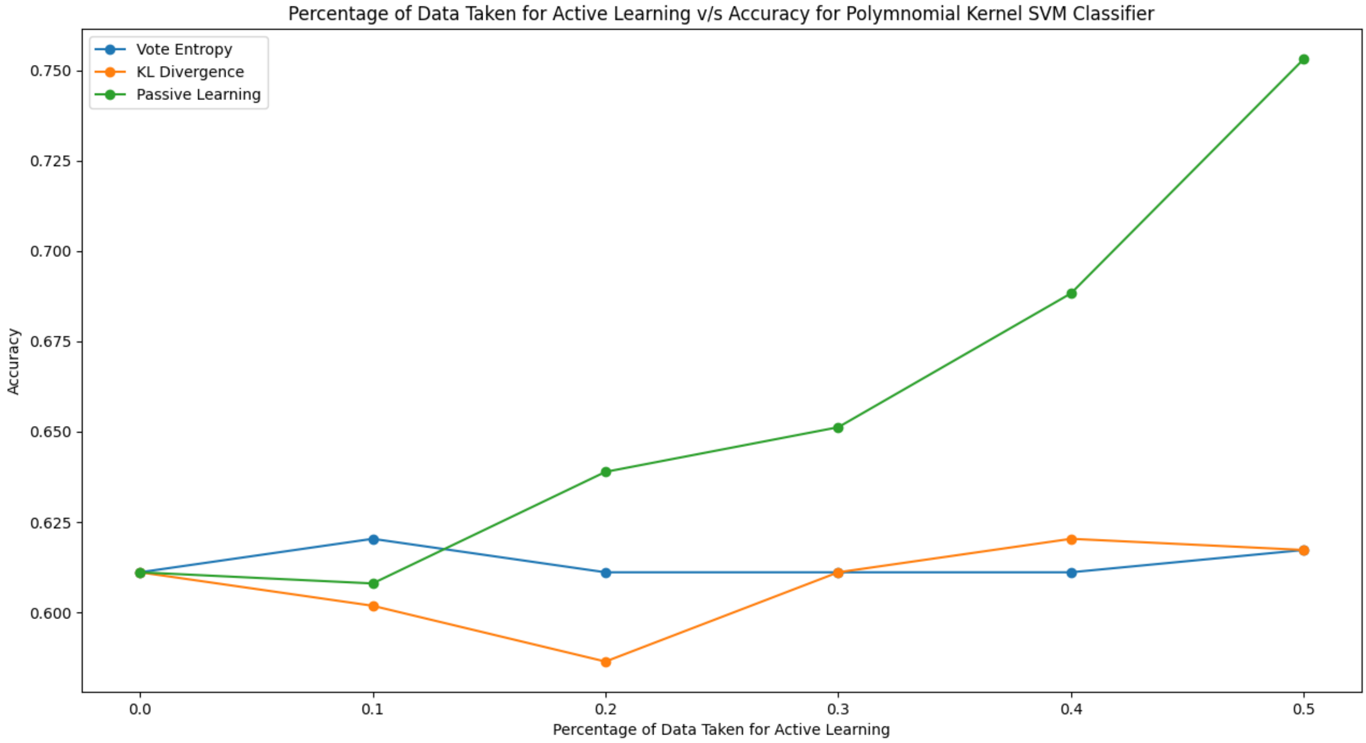
2.2 Random Forest Classifier



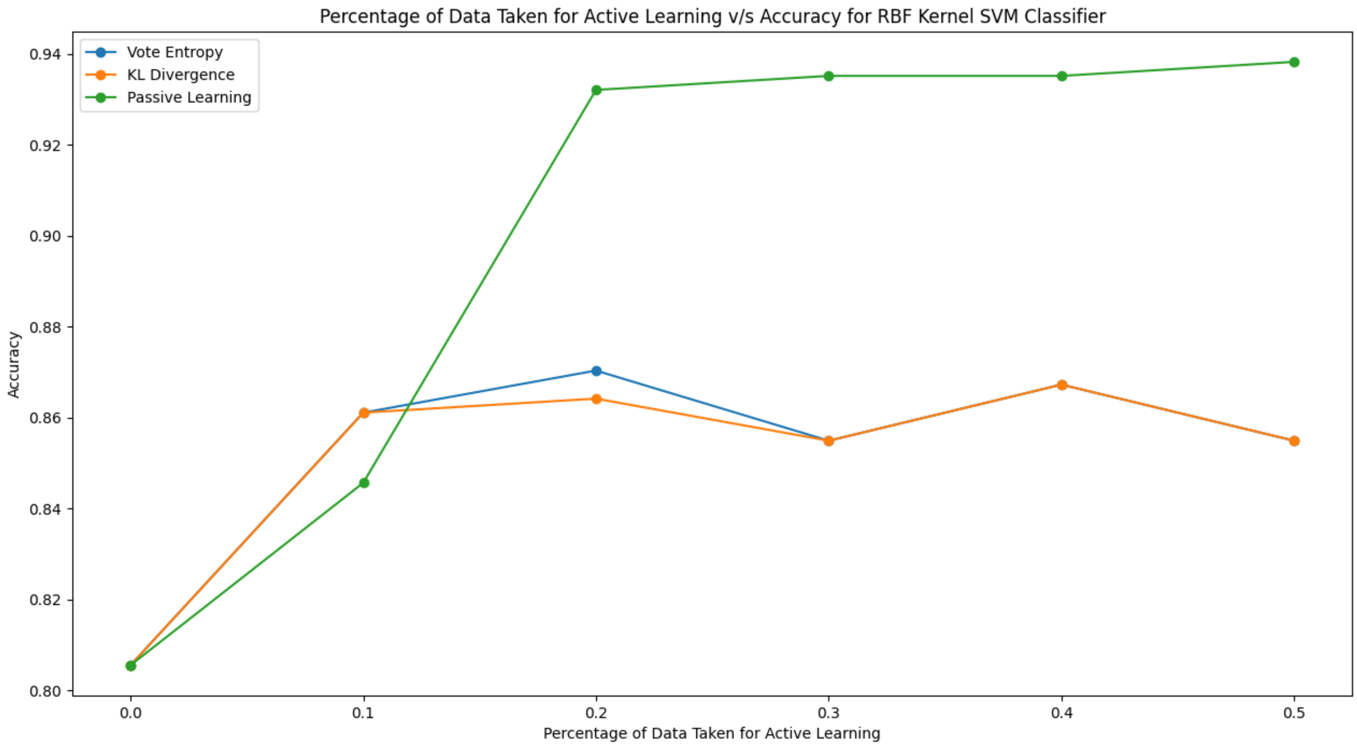
2.3 Linear SVM



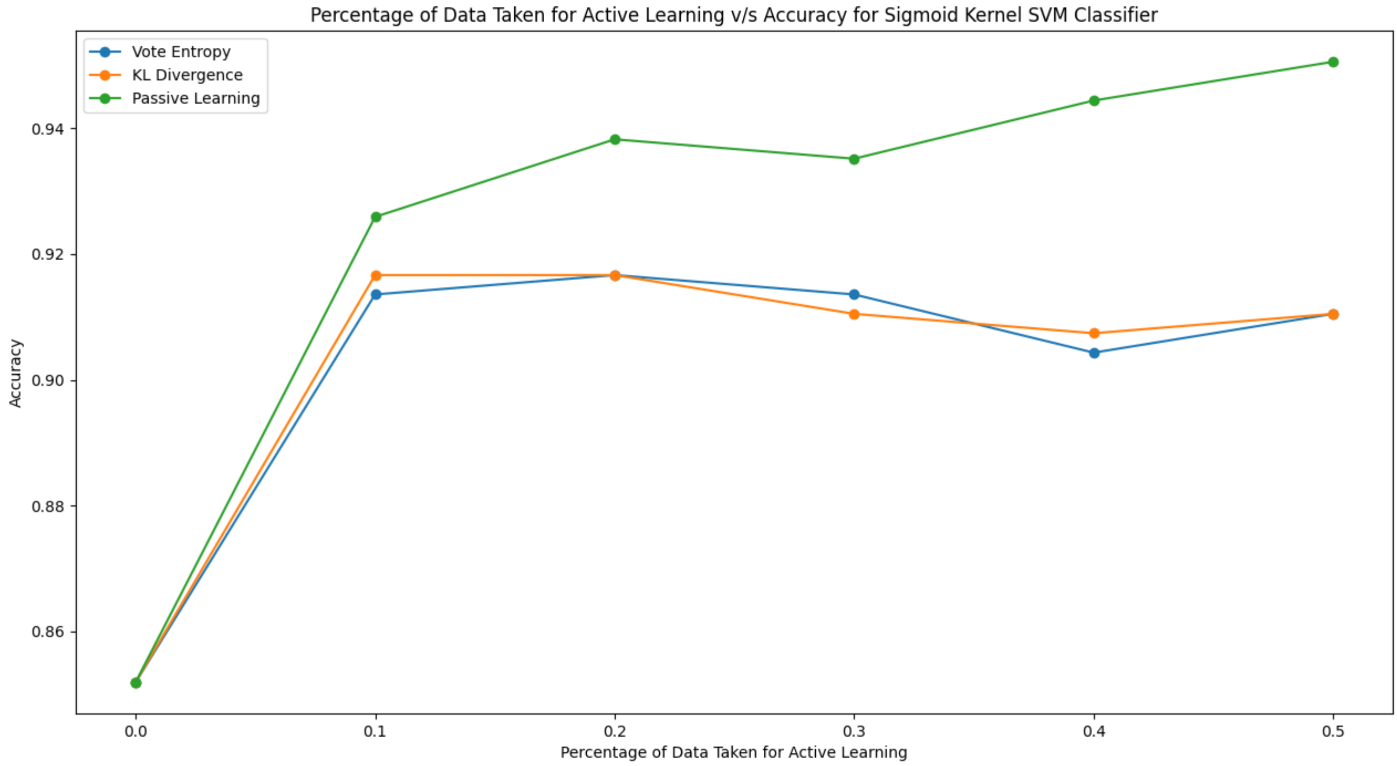
2.4 Polynomial SVM



2.5 RBF Kernel SVM

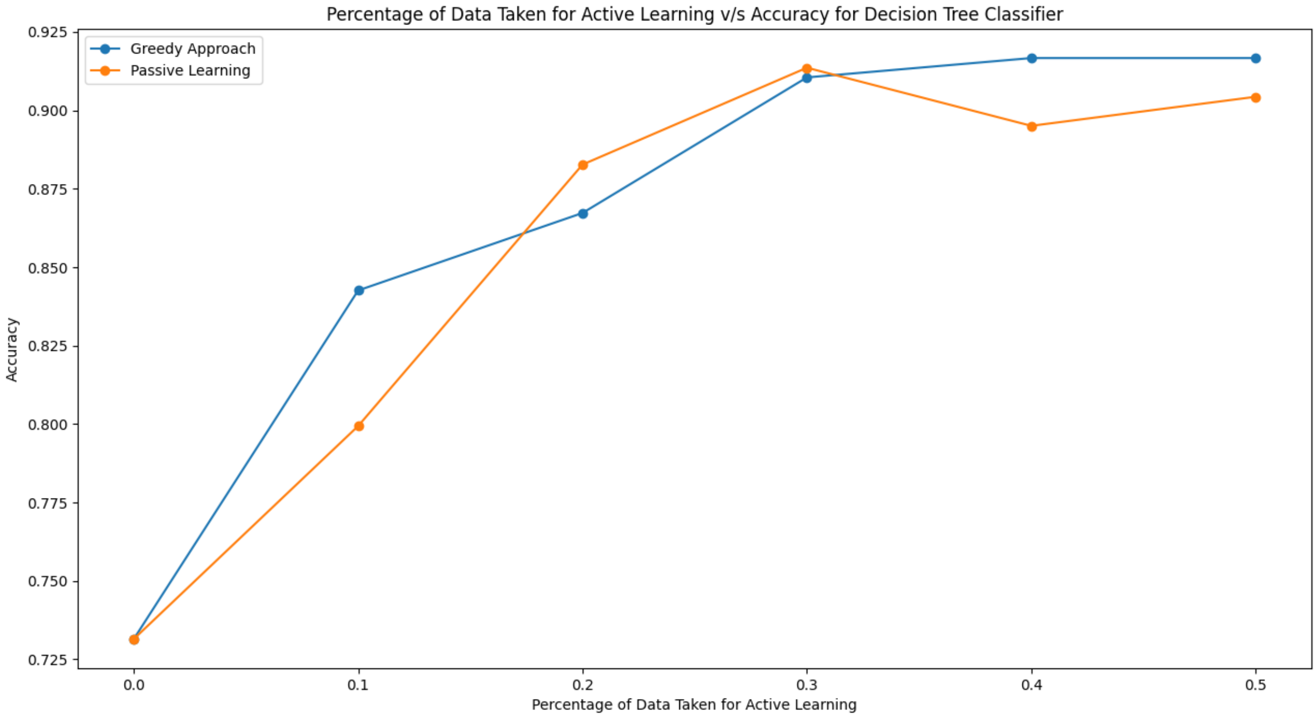


2.6 Sigmoid SVM

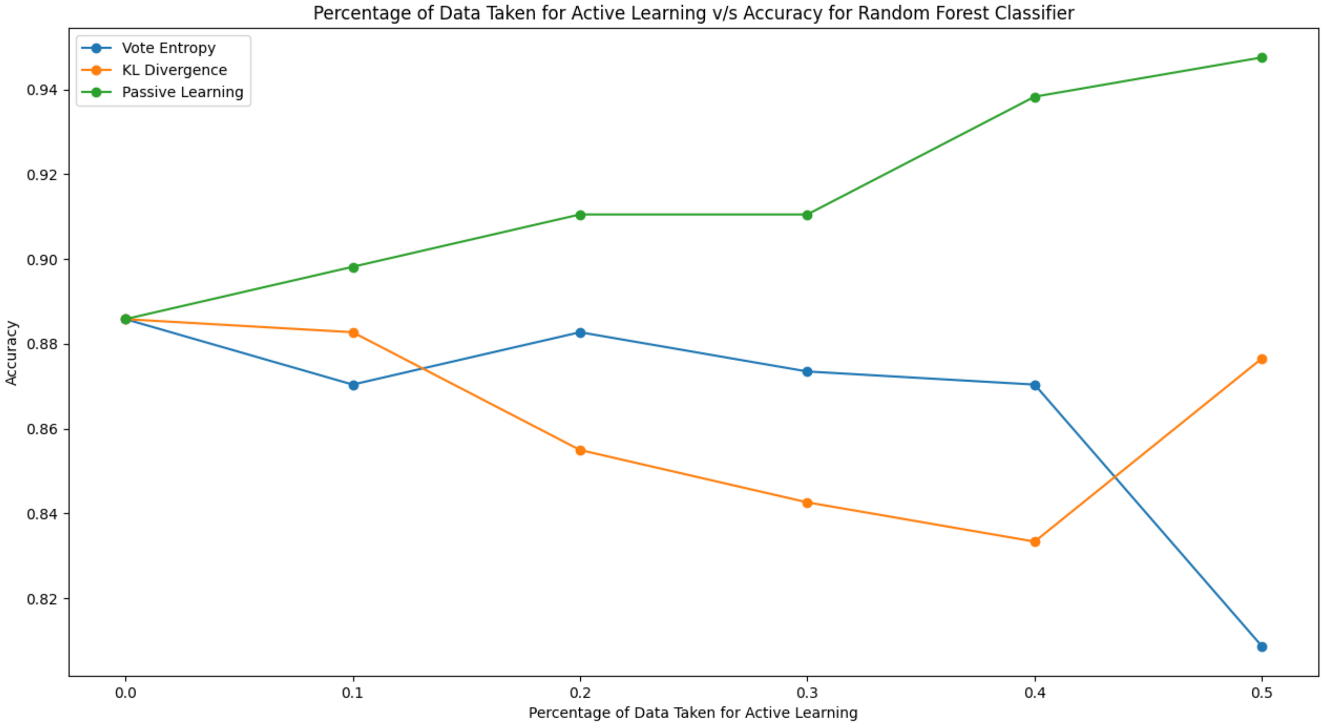


3. Greedy Approach of Reducing Version Space

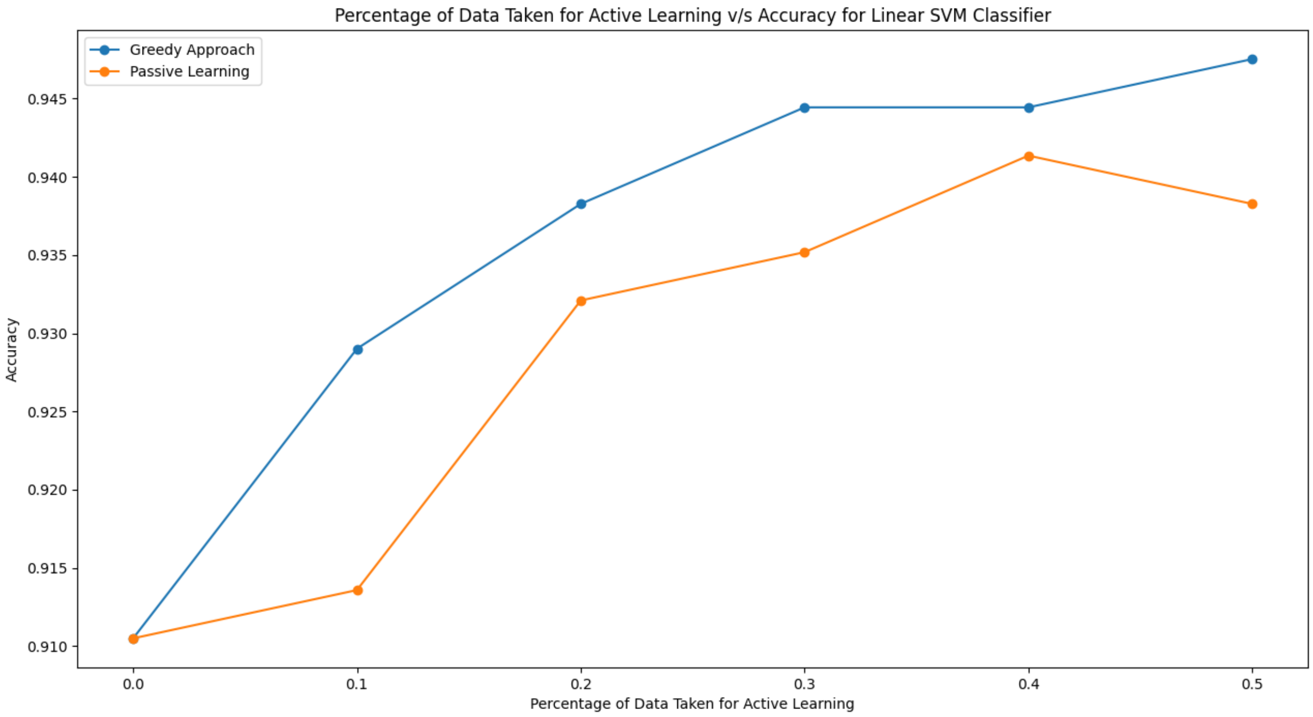
3.1 Decision Tree Classifier



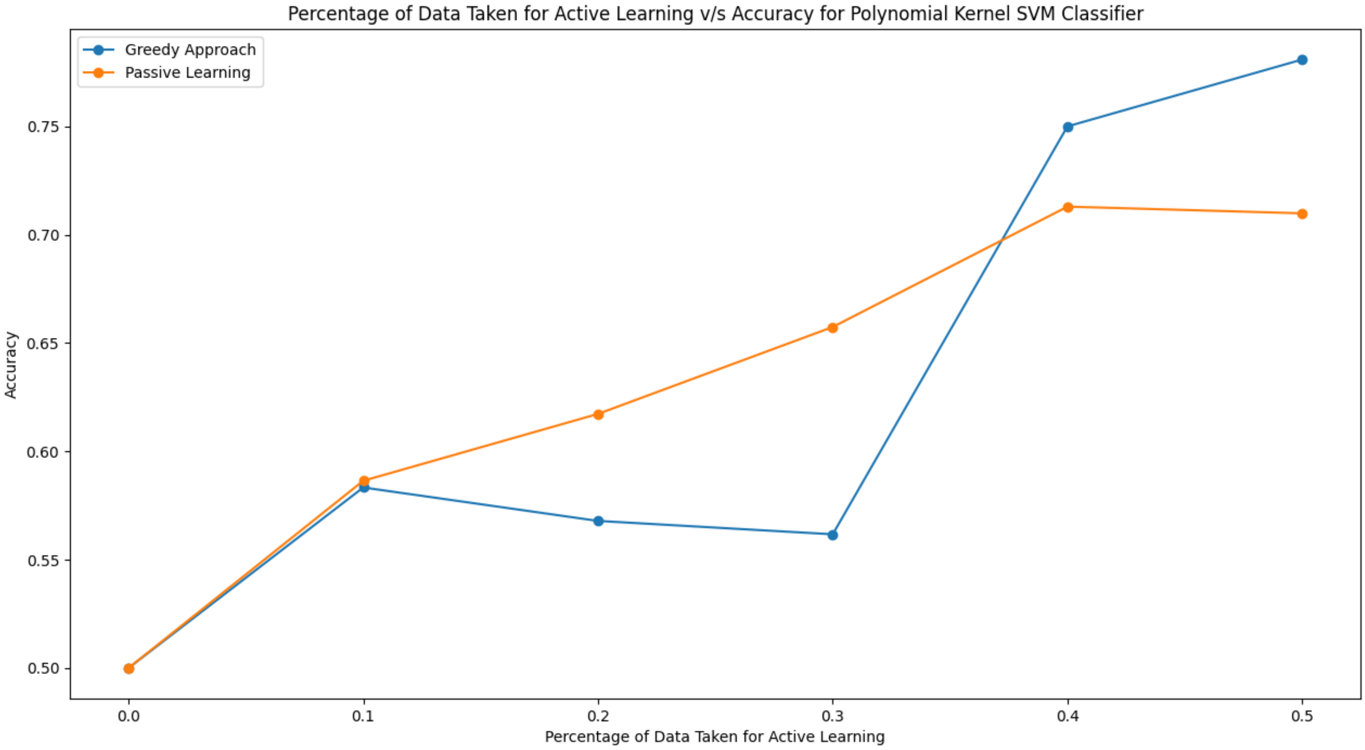
3.2 Random Forest Classifier



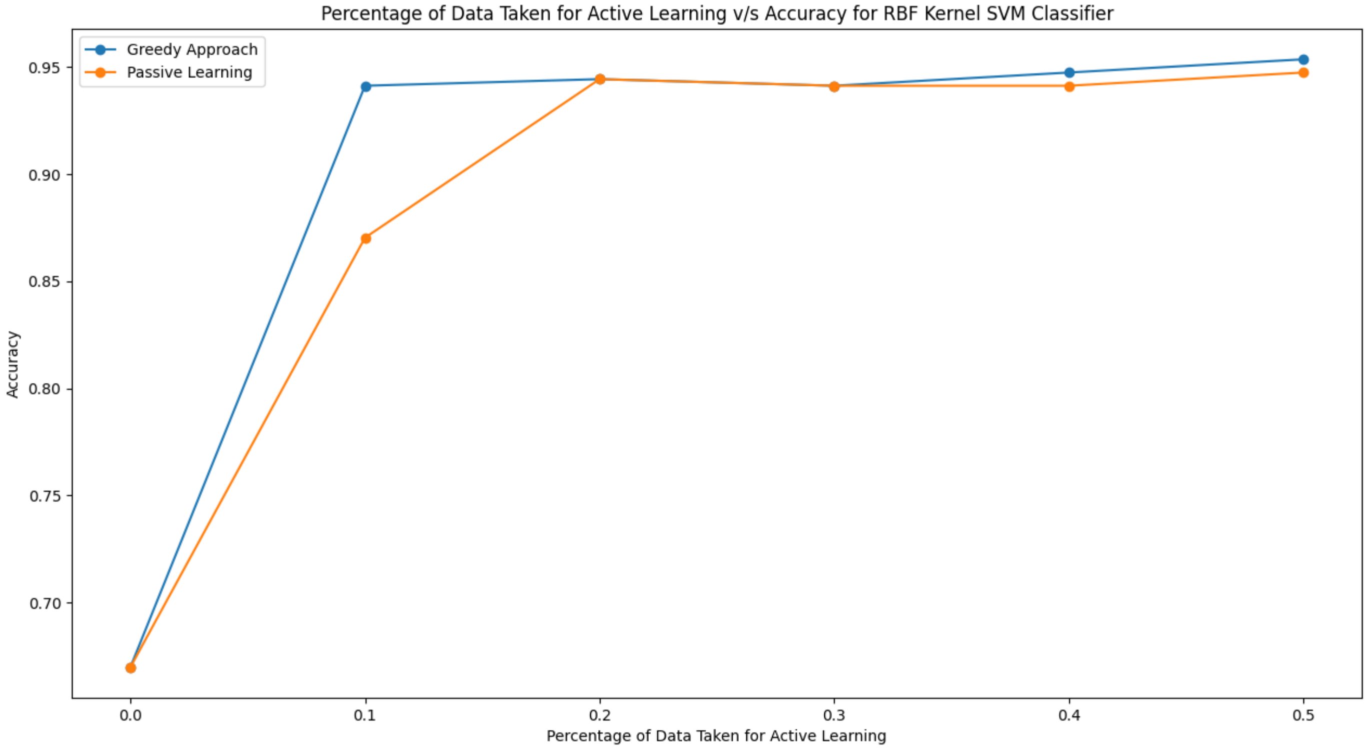
3.3 Linear SVM



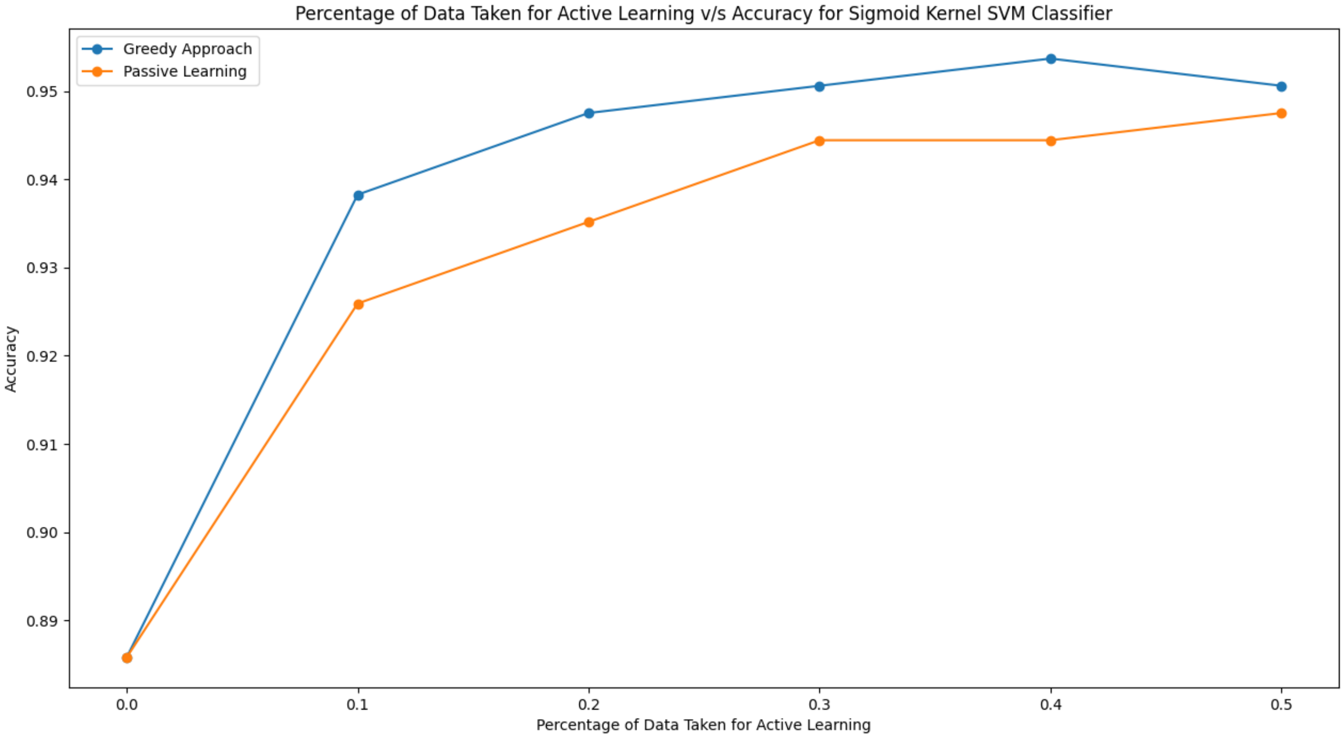
3.4 Polynomial SVM



3.5 RBF Kernel SVM

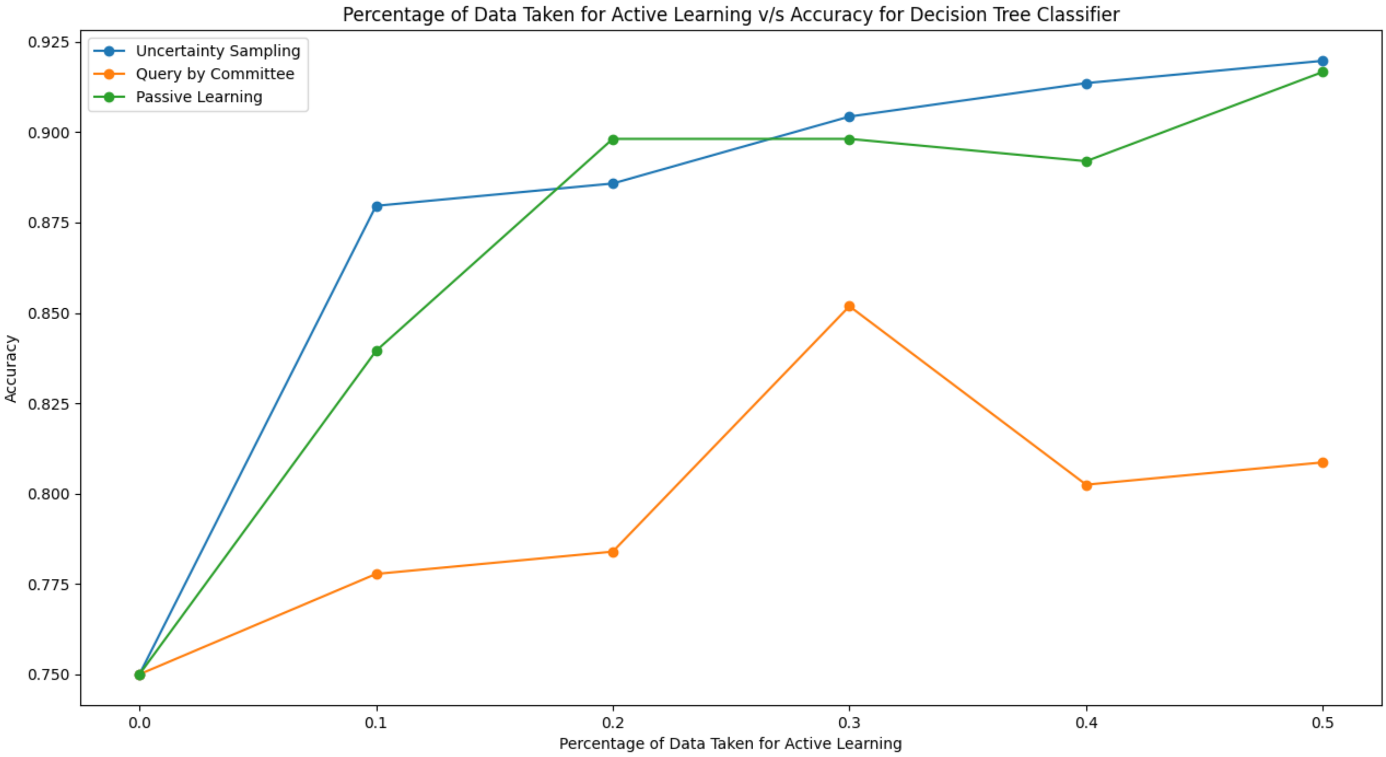


3.6 Sigmoid SVM

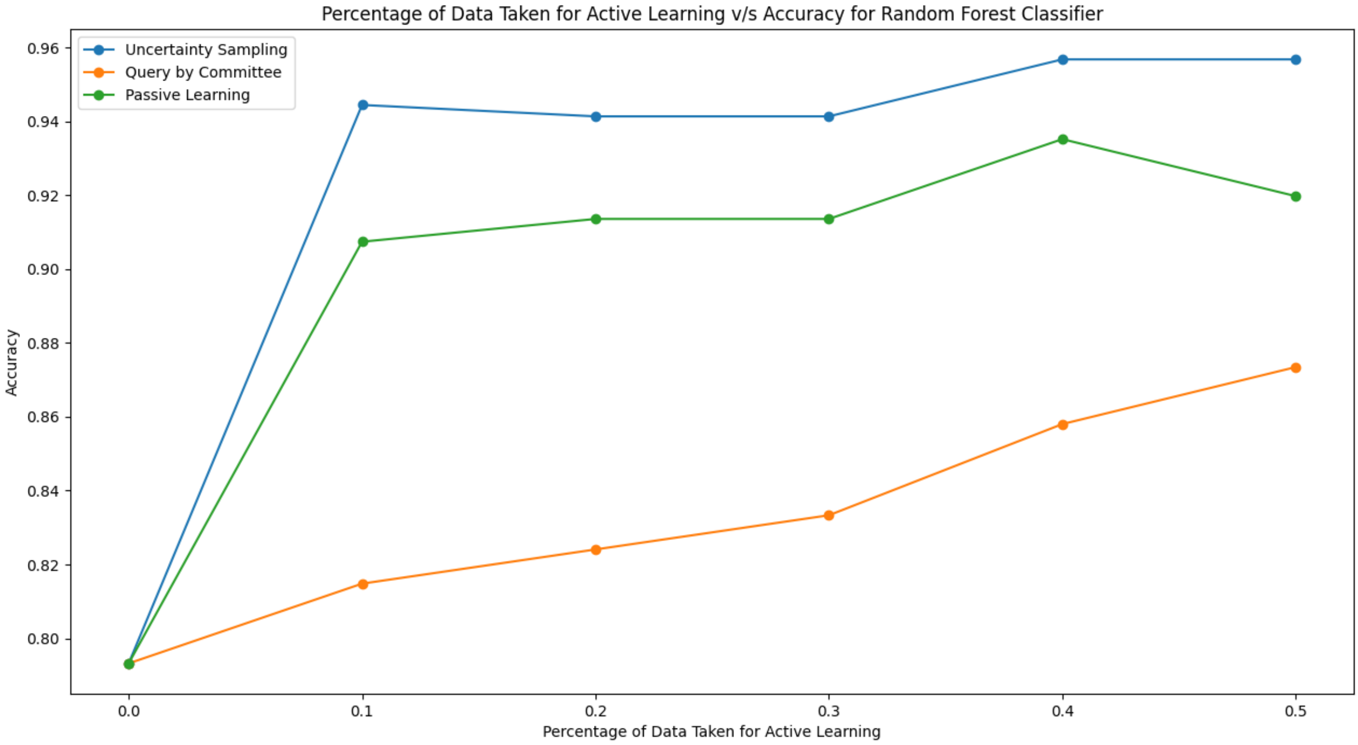


4. Combined Result of Uncertainty Sampling and QBC

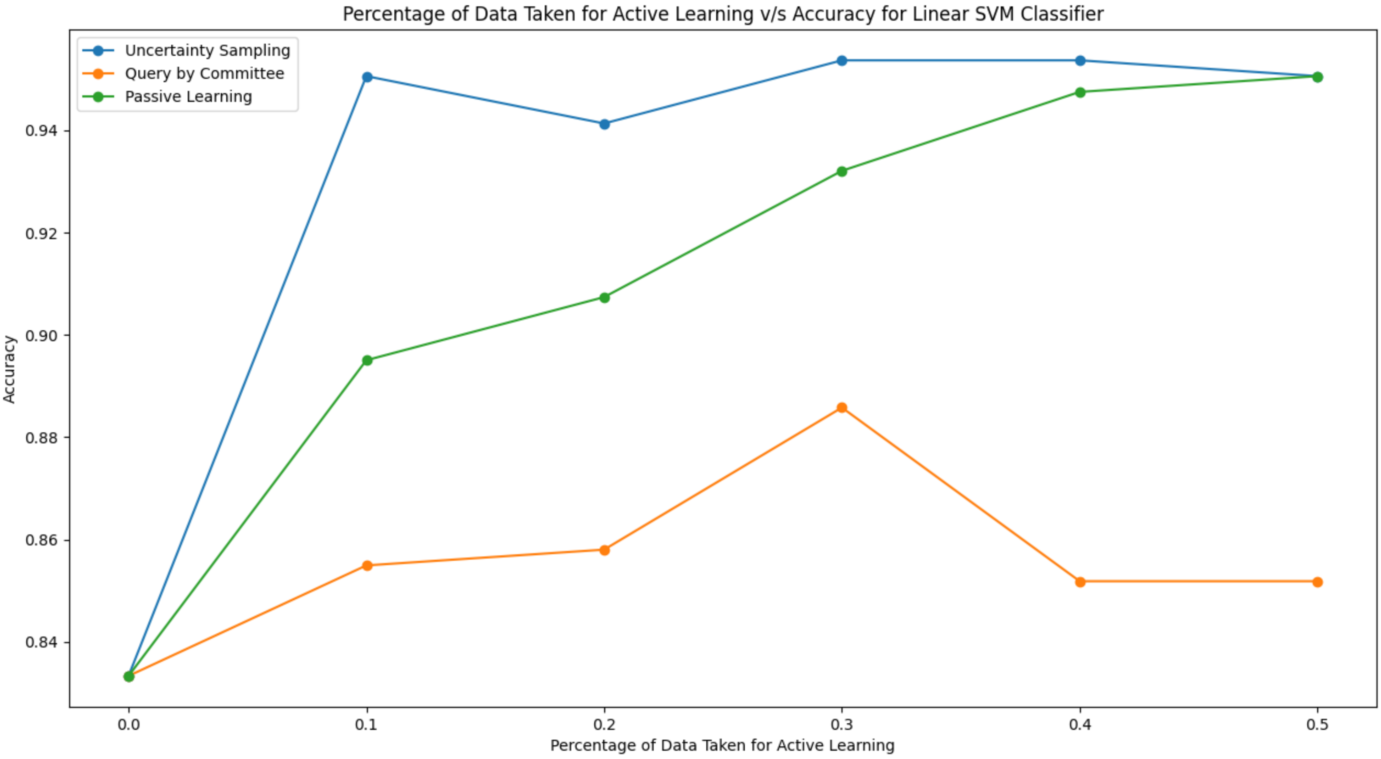
4.1 Decision Tree Classifier



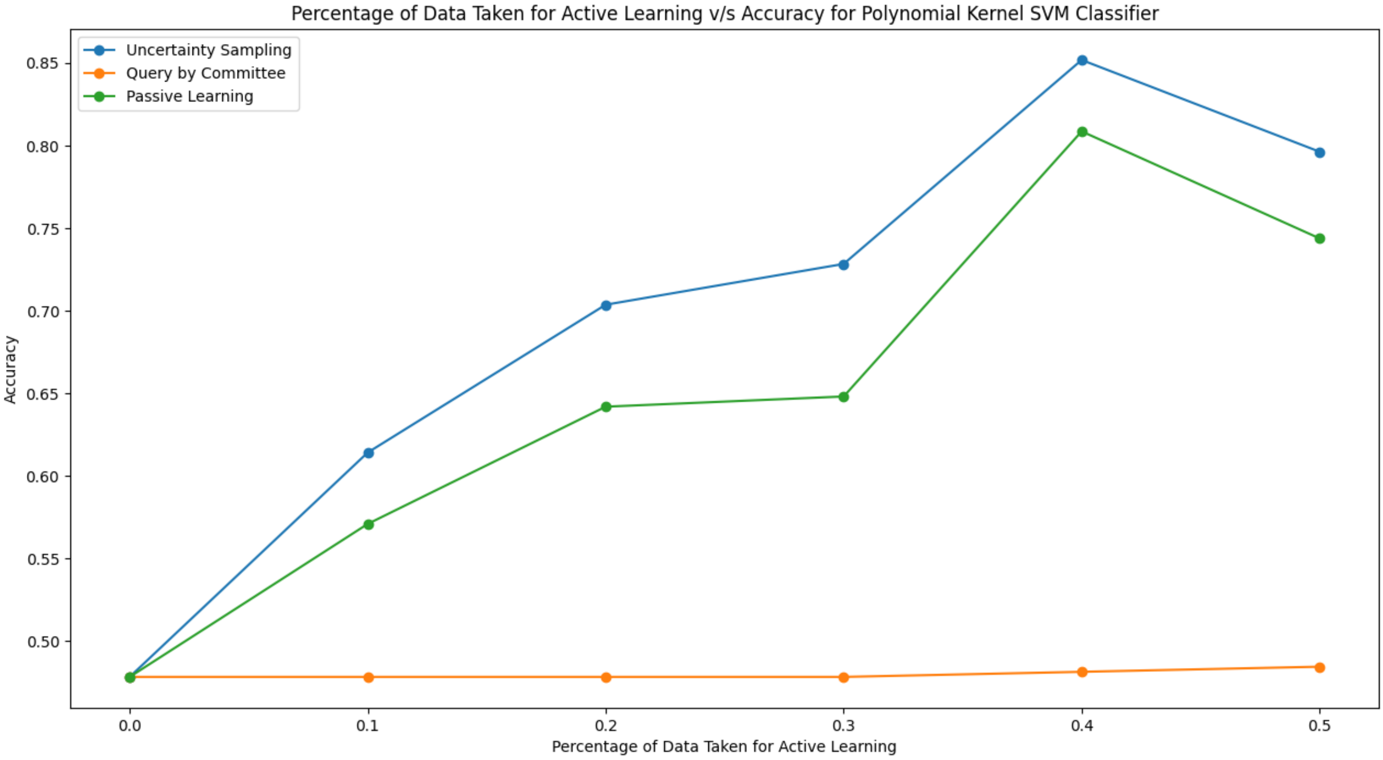
4.2 Random Forest Classifier



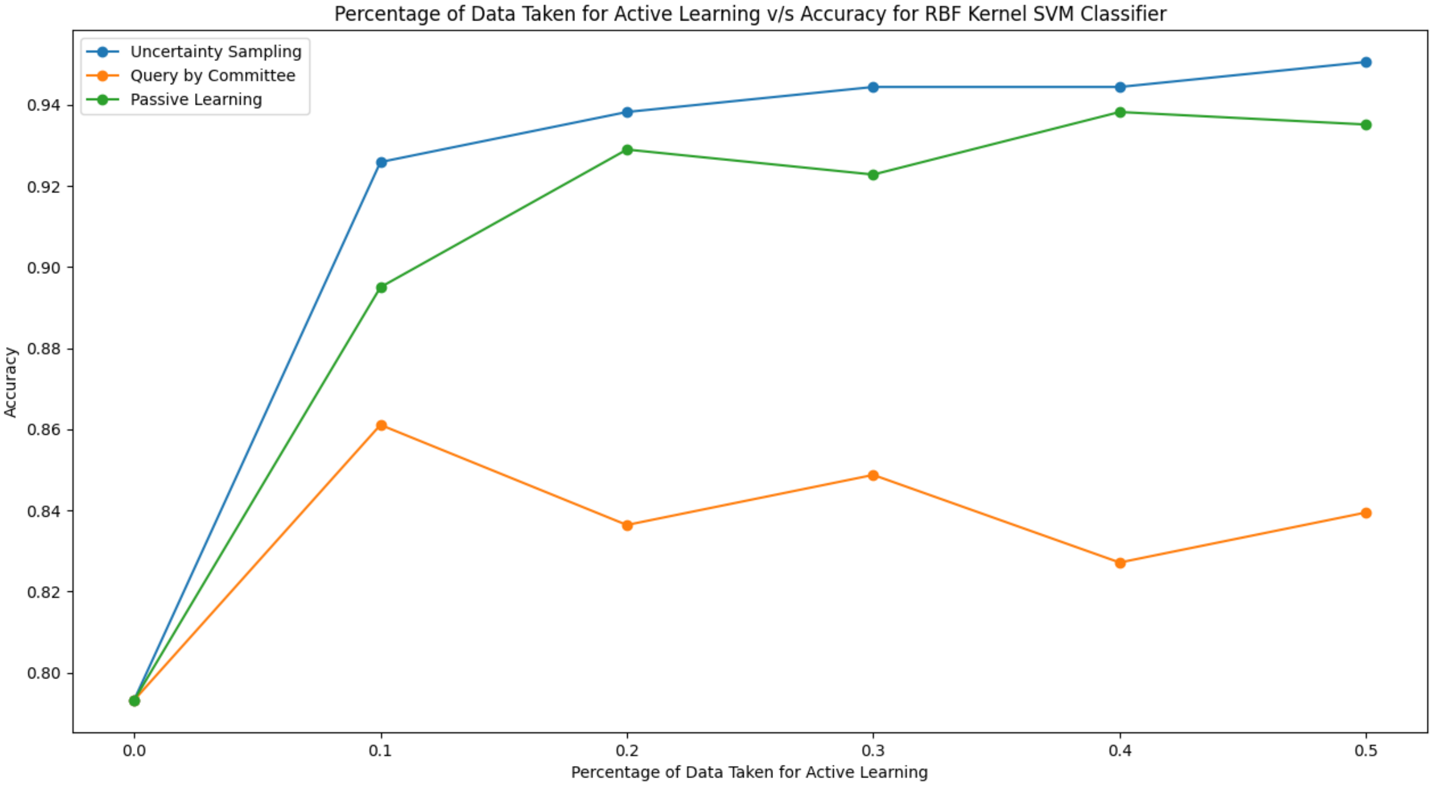
4.3 Linear SVM



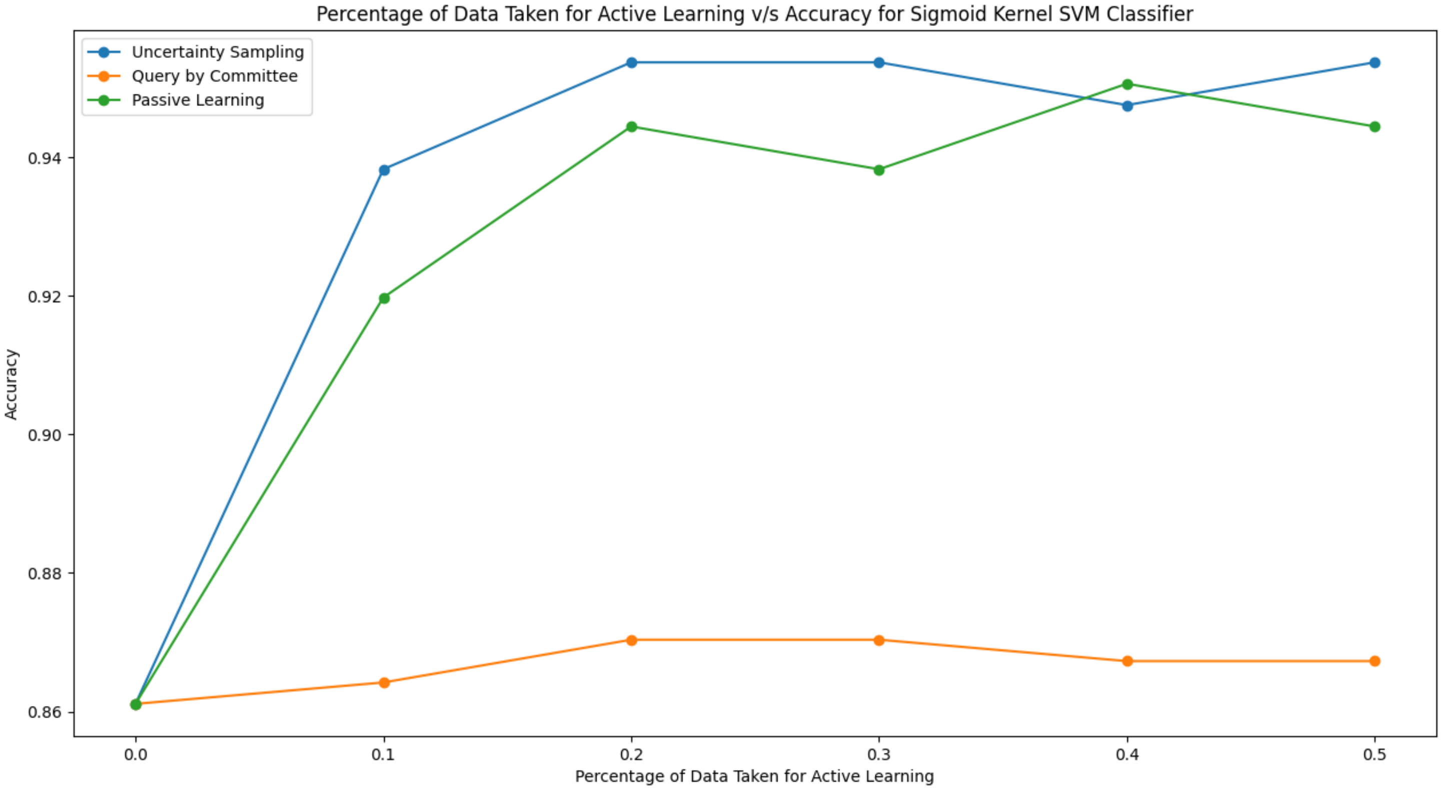
4.4 Polynomial SVM



4.5 RBF Kernel SVM



4.6 Sigmoid SVM



5. Clustering-based Classifier

Since there are 6 labels, as specified, 6 was input as K for the K-Means clustering algorithm and randomly 20% tuples were selected from each cluster and given label. The accuracy came out to be around 70%.

Stream based Active Learning for the Current Scenario

Instead of ranking the whole dataset and labelling the most uncertain ones, we can check if a tuple is belonging to version space or not. If the tuple belongs to the version space, query its label, else discard it. The working is similar to the Greedy method applied for actively labelling the dataset.