Active Learning

MULTI LABEL WINE QUALITY CLASSIFIER

REPRESENTATIVE
V SUSHANT (2019A7PS0045P)
RISHAB KABDI JAIN (2019A4PS0525P)
SHRUTI GANGWAR (2019B2A10920P)

Introduction

Active learning is learning in which a learning algorithm can interactively query a user to label new data points with the desired outputs. There are situations in which unlabeled data is abundant but manual labeling is expensive. Active learning can be used in situations where the data is too large to be labelled and priority needs to be made to label the data in a smart way.

About Dataset

The dataset is related to the red variant of the Portuguese "Vinho Verde" wine. It is based on 11 parameters- fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol values . Then, a quality score is decided in between 3 to 8. This dataset was taken from UCI Machine Learning Repository.

modAL

modAL is the library that is majorly used to compute the Active Learning tasks in our code. Following which the data is split into Labelled data, unlabelled data, Labels and Oracle answers in a 90:10 split.

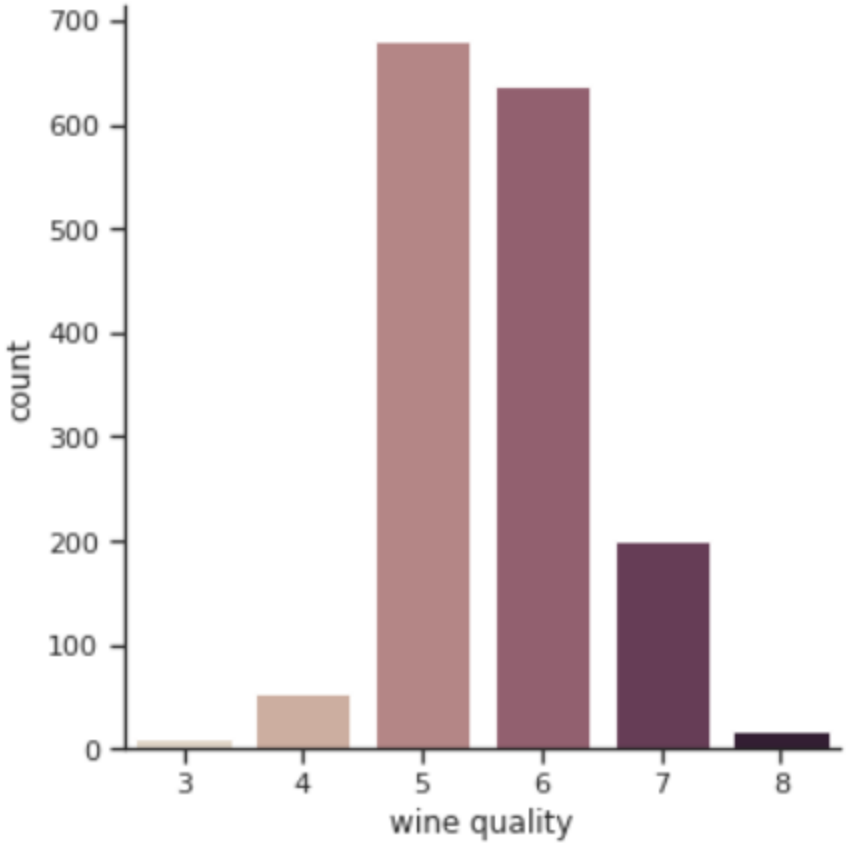
Following which using Active Learner an instance is made with Random Forest Classifier as an estimator and Uncertainty sampling (Least confidence Interval) as query strategy,

where our X is the Labelled data and Y is labels. Here, Active learner queries instances about which it is least certain how to label,thus the data is sampled through uncertainty sampling (Least confident sampling).

Following which 4 sets of variables are made where the unlabelled points are split into additional 10%, 20%, 30% and 40% points respectively. These sets of variables are converted into numpy arrays.

Exploratory Analysis & Data Visualizations

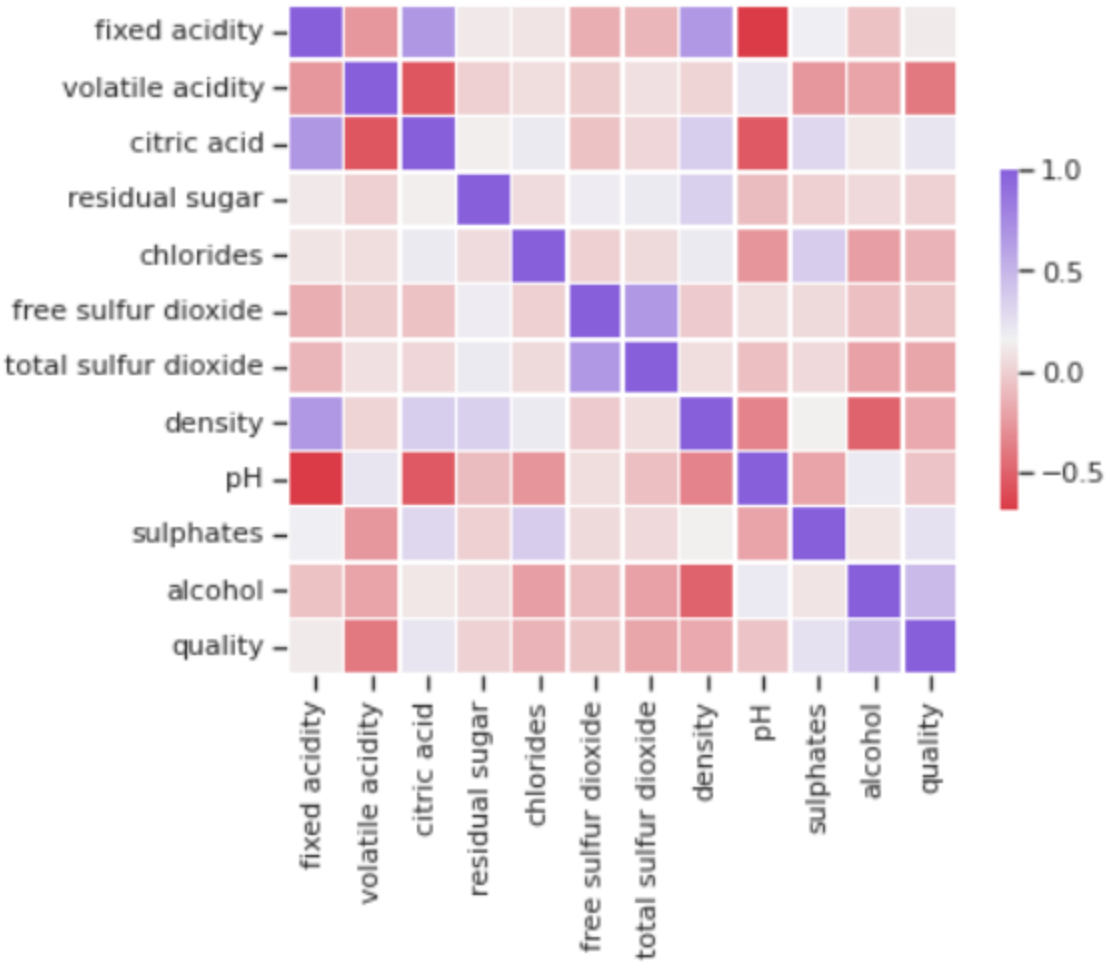
A histogram between count vs quality is plotted to understand the skewness of the data. The dataset is extremely skewed with minority classes like '3' and '8' sharing less than 1% of the total population.



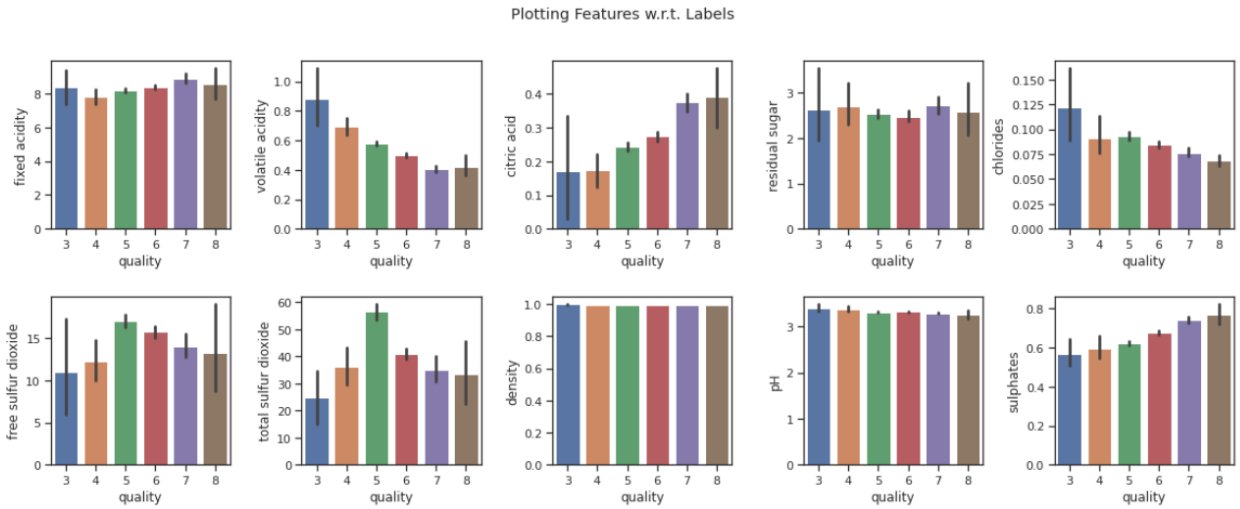
We can see this by plotting a histogram on the 'quality' column.

Correlation Analysis

To understand the correlation between different variables of the dataset, a heatmap is visualised.



Further visualizing the relations between features and wine quality, we have

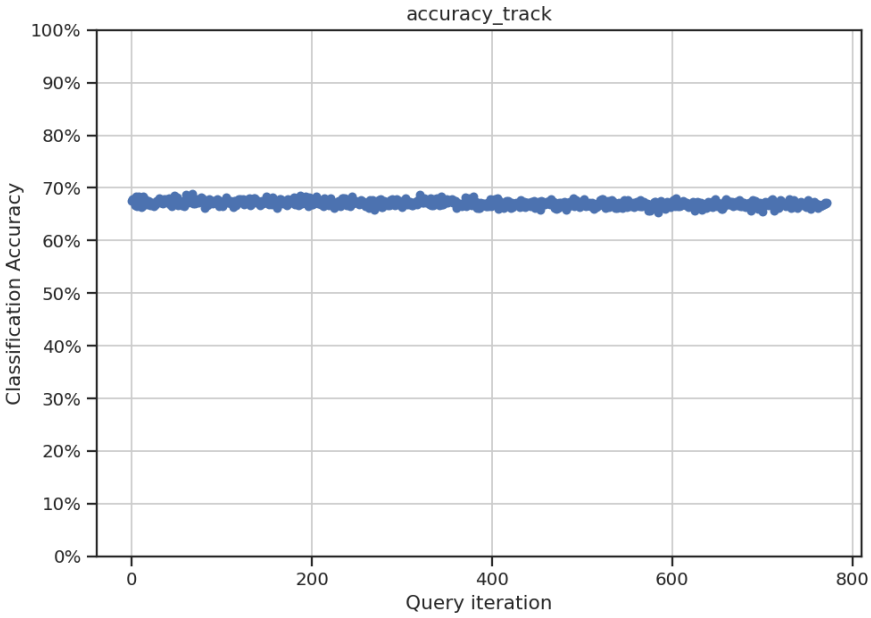


Data Preprocessing

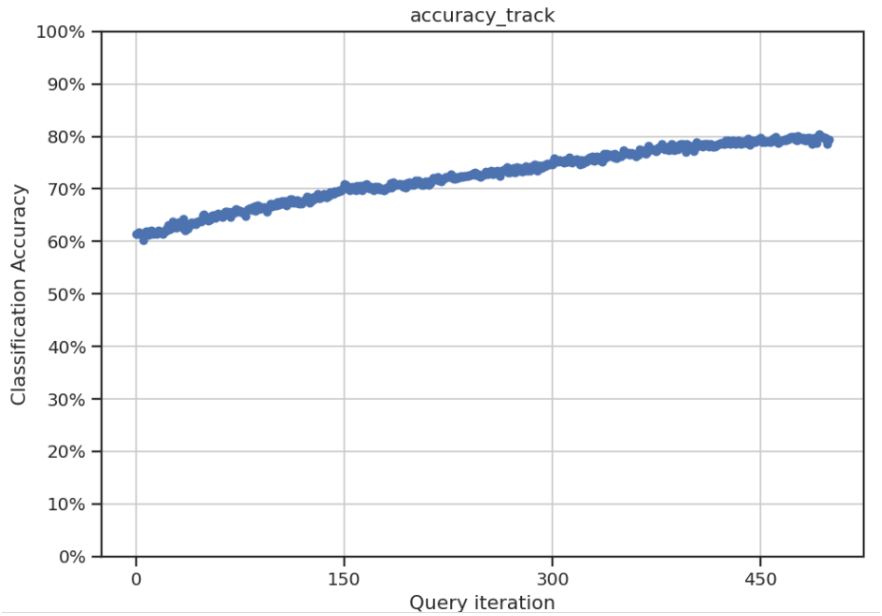
The dataset that we have when divided into X and Y variables, X has a shape of (1599, 11), and Y has a shape of (1599). The data is complete and does not contain any null or missing values.

Uncertainty Sampling with Least confidence

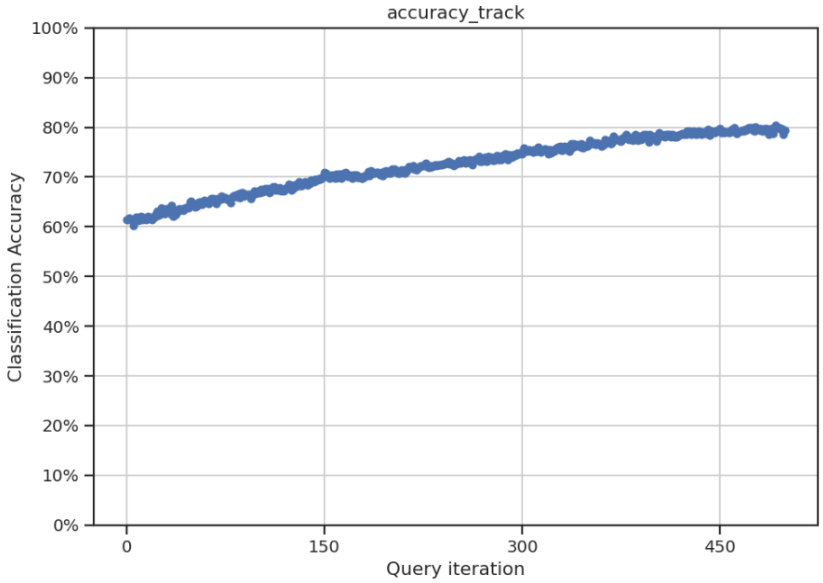
The first set of variable with 10% split is fit into Active Learner with Uncertainty sampling with multiple queries.Required post-processing is done of the data after oracle gives the input.The accuracy vs no. of queries graph is plotted-



Similar process is done for the 30% split data.The accuracy vs no. of queries graph is plotted-

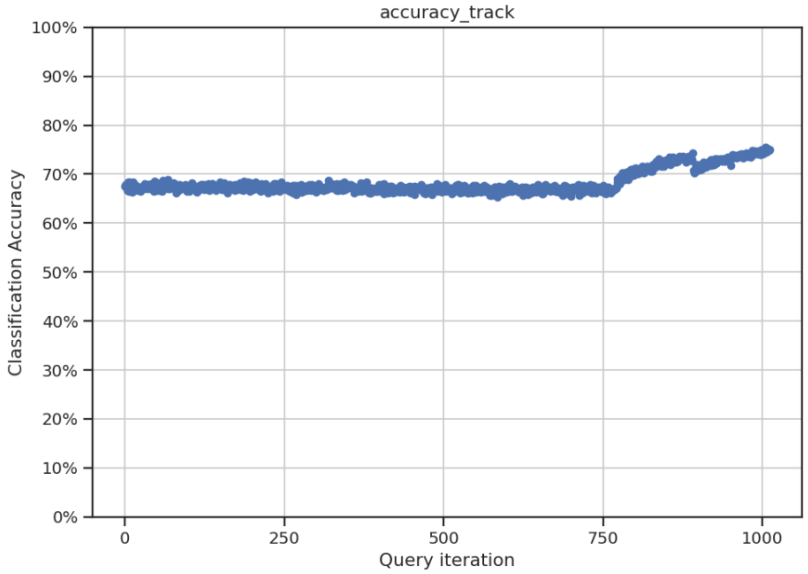


Similar process is done for the 40% split data.The accuracy vs no. of queries graph is plotted-

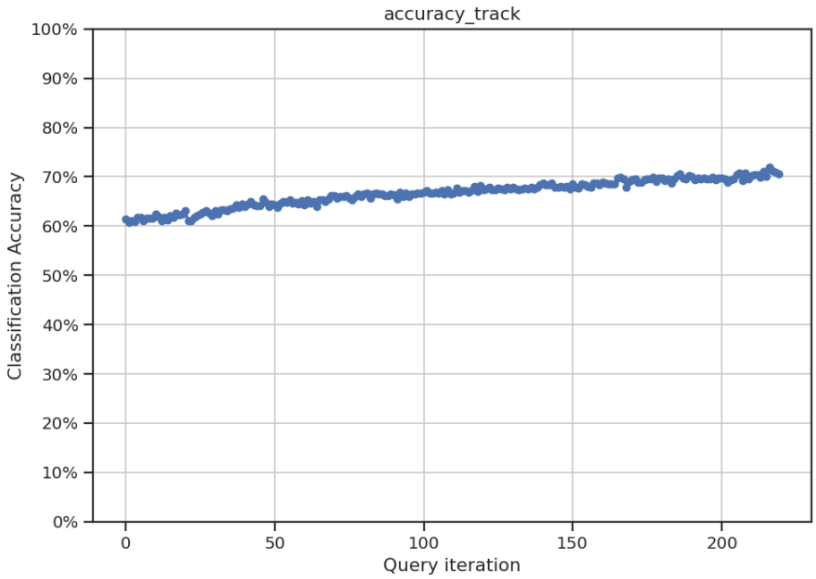


Minimum Margin Sampling

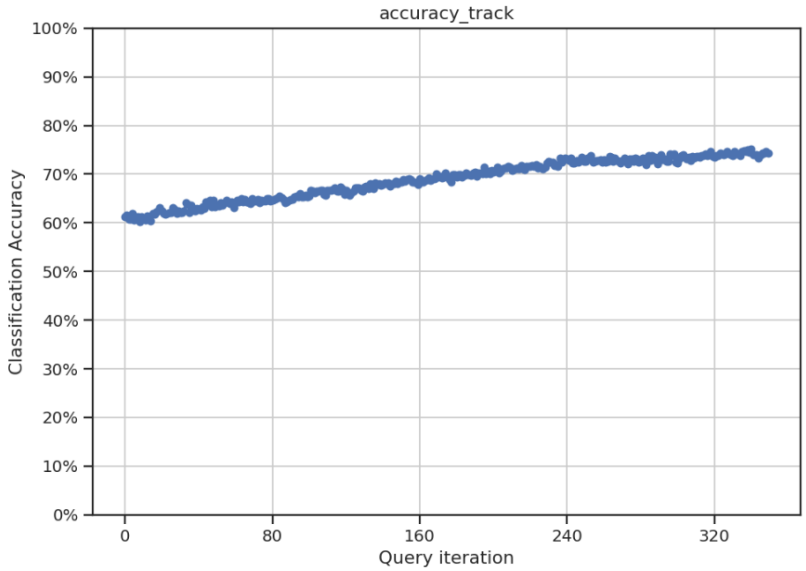
The first set of variable with 10% split is again fit into Active Learner with Least Margin sampling with multiple queries.Required post-processing is done of the data after oracle gives the input.The accuracy vs no. of queries graph is plotted-



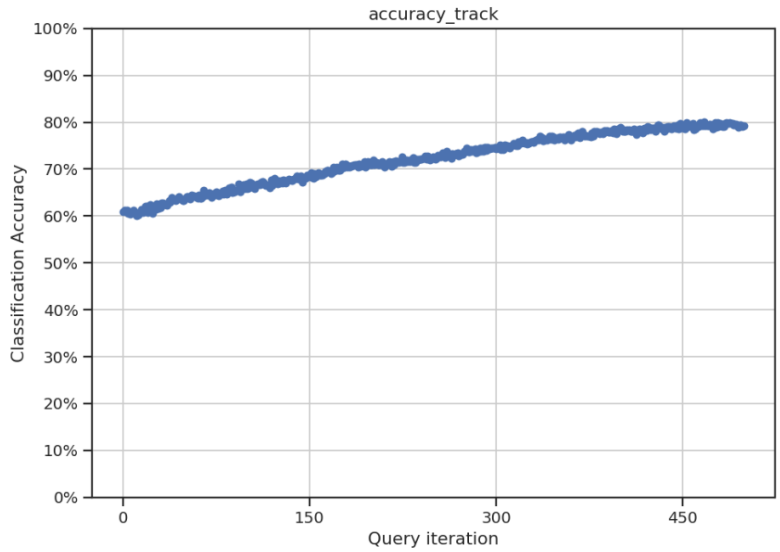
Similar process is done for the 20% split data.The accuracy vs no. of queries graph is plotted-



Similar process is done for the 30% split data.The accuracy vs no. of queries graph is plotted-

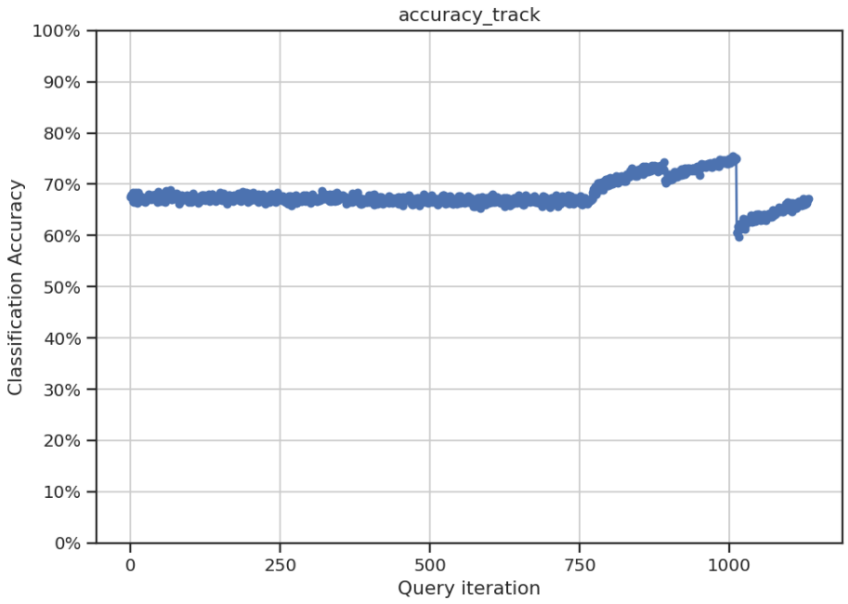


Similar process is done for the 40% split data.The accuracy vs no. of queries graph is plotted-

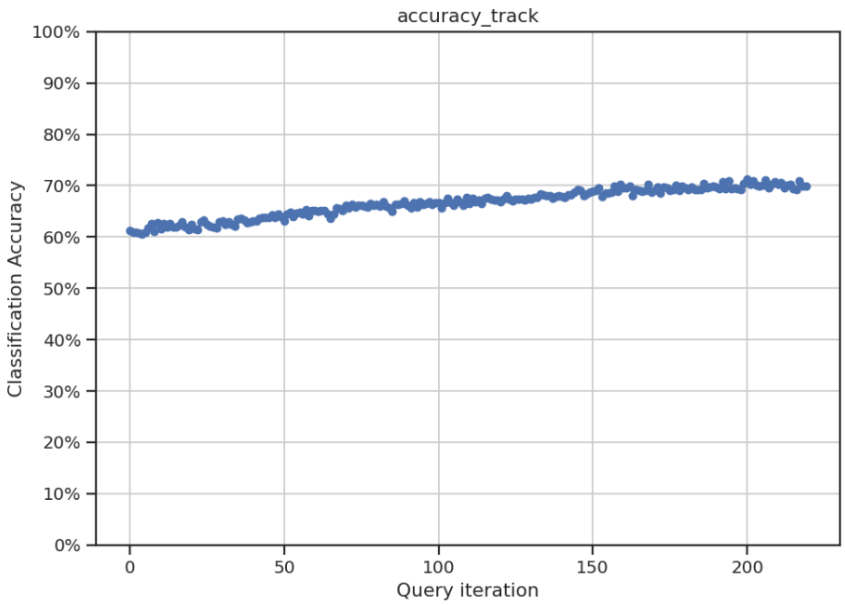


Entropy Sampling

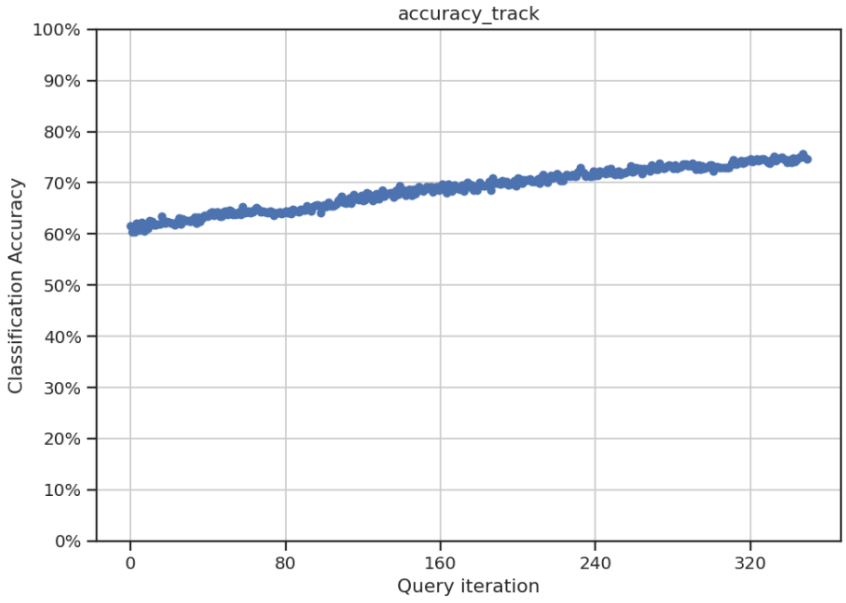
The first set of variable with 10% split is again fit into Active Learner with Entropy sampling with multiple queries.Required post-processing is done of the data after oracle gives the input.The accuracy vs no. of queries graph is plotted-



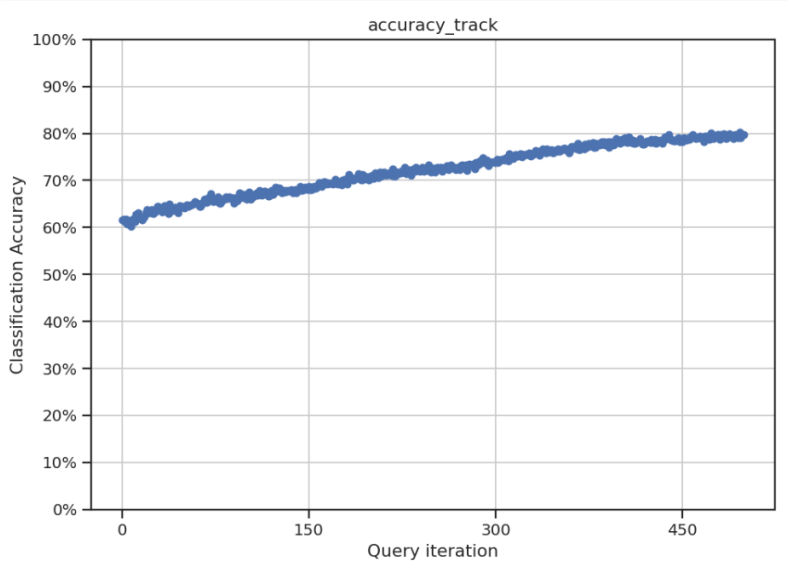
Similar process is done for the 20% split data.The accuracy vs no. of queries graph is plotted-



Similar process is done for the 30% split data.The accuracy vs no. of queries graph is plotted-

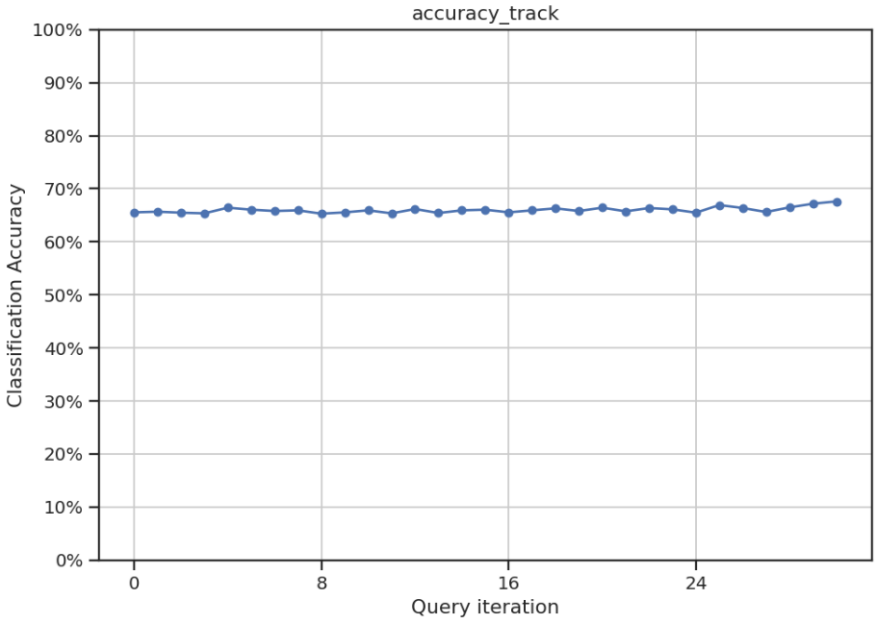


Similar process is done for the 40% split data.The accuracy vs no. of queries graph is plotted-



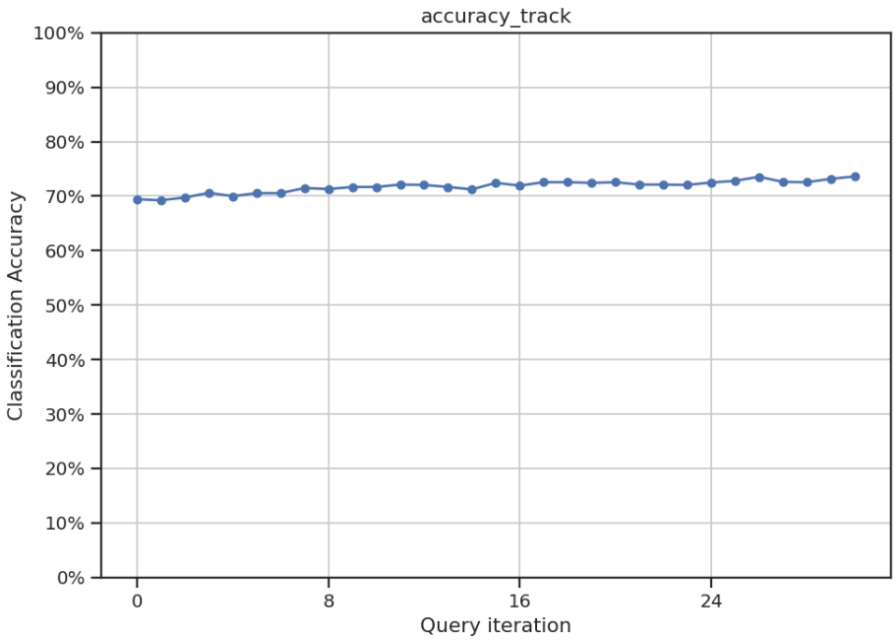
Vote Entropy

5 different committee members are made and are fit into Active Learner where points are labelled using Vote Entropy Method.The accuracy vs no. of queries graph is plotted-



KL DIVERGENCE

5 different committee members are made and fit into Active Learner where points are labelled using KL Divergence Method. The accuracy vs no. of queries graph is plotted-



For incorporating Stream based learning in our problem, firstly we need to select each instance one by one randomly from an entire pool of Mislabelled data for classifier.For uncertainty greater than 0.5 ,the Oracle is used for querying and then added to the label data pool.

References

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.
   In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
2. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
3. https://readthedocs.org/projects/modal-python/downloads/pdf/latest/
   https://github.com/yixin0829/multi-label-wine-quality-classification/blob/master/wine_quality_multi_classification_final.ipynb

4. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.
   In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
5. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
6. https://readthedocs.org/projects/modal-python/downloads/pdf/latest/
   https://github.com/yixin0829/multi-label-wine-quality-classification/blob/master/wine_quality_multi_classification_final.ipynb