Analyze different active learning strategies with scikit-activeml

20200031	Ahmed Ali Ali
20210597	Kareem Sayed Salah
20200347	Omar Salah eldin Ahmed
20200807	Ahmed Ashraf Zaky

1. Introduction

Active learning is a machine learning paradigm that aims to iteratively select the most informative data points for labeling, thus reducing the need for extensive labeled data and potentially improving model performance. In this report, we explore the application of active learning strategies using the scikit-activeml library in two distinct datasets: Balanced dataset (the Loan Approval Dataset) and Imbalanced dataset (the Diabetes Prediction Dataset).

2. Datasets

2.1. Balanced Dataset

Loan Approval Dataset

The Loan Approval Prediction Dataset comprises financial records and associated information utilized to assess the eligibility of individuals or organizations for obtaining loans from lending institutions. This dataset is commonly used in machine learning and data analysis to develop predictive models and algorithms that determine the likelihood of loan approval based on various factors.

Dataset Characteristics:

The dataset contains 3296 sample in total divided into:

• 2966 training set

• 330 test set

Categorical Columns:

1. **Gender**: Categorical variable indicating the gender of the applicant (Male/Female).
2. **Married**: Binary variable indicating whether the applicant is married (Yes/No).
3. **Number of Dependents**: Categorical variable representing the number of dependents (Possible values: 0, 1, 2, 3+).
4. **Education**: Categorical variable indicating the educational qualification of the applicant (Graduate/Not Graduate).
5. **Self-Employed**: Binary variable indicating whether the applicant is self-employed (Yes/No).
6. **Credit History**: Binary variable indicating the credit history of the applicant (Yes/No).
7. **Property Area**: Categorical variable indicating the location of the property (Rural/Semi-Urban/Urban).
8. **Loan Status**: Target variable indicating loan approval status (Y/N).

Numerical Columns:

1. **Loan ID**: Unique identifier for each loan application.
2. **Applicant Income**: Numerical variable representing the income of the primary applicant.
3. **Co-applicant Income**: Numerical variable representing the income of the co-applicant (if applicable).
4. **Loan Amount**: Numerical variable representing the requested loan amount.
5. **Loan Amount Term**: Numerical variable representing the term (in months) of the loan amount.

Target Variable:

- **Loan Status**: Binary variable indicating whether the loan was approved (Y) or not (N).

2.2. Imbalanced Dataset

Diabetes Prediction Dataset Description

The Diabetes Prediction Dataset consists of data related to individuals' health parameters and medical history, primarily aimed at predicting the likelihood of diabetes occurrence. This dataset is designed to facilitate the development of predictive models and algorithms for identifying individuals at risk of diabetes based on their health attributes.

Dataset Characteristics:

The dataset contains 100000 sample in total divided into:

• 67000 training set

• 33000 test set

Numeric Variables:

1. Age
2. BMI (Body Mass Index)
3. HbA1c Level
4. Blood Glucose Level

Categorical Variables:

1. Gender
2. Smoking History (Assuming categories like "Never smoked", "Former smoker", "Current smoker", etc.)

Binary Variables:

1. Hypertension
2. Heart Disease

Target Variable:

1. Diabetes (Binary variable indicating the presence of diabetes - 1 for yes, 0 for no

3. Active Learning Strategy

We apply active learning using uncertainty sampling, margin sampling, entropy sampling and Query by Committee (KL Divergence and Vote Entropy) methods to improve loan approval prediction models.

4. Methodology

We use a K-Nearest Neighbors classifier and evaluate model performance over 20 iterations, updating the model with newly labeled data points in each iteration.

5. Comparison of Active Learning Strategies

1. Uncertainty Sampling:
    - Principle: Prioritizes samples where the model is most uncertain about predictions.
    - Focus: Instances near the decision boundary.
    - Objective: To label data points that are expected to provide the most information to the model.
    - Strengths: Effective in scenarios where decision boundaries are well-defined.
    - Limitations: May miss informative instances that are far from the decision boundary.

2. Entropy Sampling:

- Principle: Selects instances with the highest information entropy, indicating higher uncertainty.
- Focus: Maximize the reduction in uncertainty in the model's predictions.
- Objective: To reduce overall uncertainty in the model.
- Strengths: Can capture instances with diverse and complex decision boundaries.
- Limitations: May not always prioritize instances with the most significant impact on model performance.

3. Margin Sampling:

- Principle: Identifies instances with the smallest margins between the top two predicted classes.
- Focus: Samples where the model's confidence is lowest.
- Objective: Capture informative instances at the decision boundary.
- Strengths: Effective in scenarios where classes are well-separated.
- Limitations: May not perform well in scenarios with overlapping classes.

Query by Committee:

4. KL Divergence:

- Principle: Utilizes a committee of models to select instances where there is significant disagreement in predictions measured by KL divergence.
- Focus: Instances where the models in the committee disagree the most.
- Objective: Selects instances that are not only uncertain but also informative in improving the model's performance.
- Strengths: Exploits model diversity and captures informative instances even in complex scenarios.

- Limitations: Requires maintaining multiple models, potentially increasing computational complexity.

5. Vote Entropy:

- Principle: Utilizes a committee of models to select instances where there is high entropy in predictions.
- Focus: Instances where there is high uncertainty among the models' predictions.
- Objective: Selects instances that are uncertain and potentially informative in improving the model's performance.
- Strengths: Captures uncertainty and disagreement among models, providing a more robust selection criterion.
- Limitations: Similar to KL Divergence, requires maintaining multiple models, potentially increasing computational complexity.

6. Results

6.1. Loan Approval Dataset Experiment

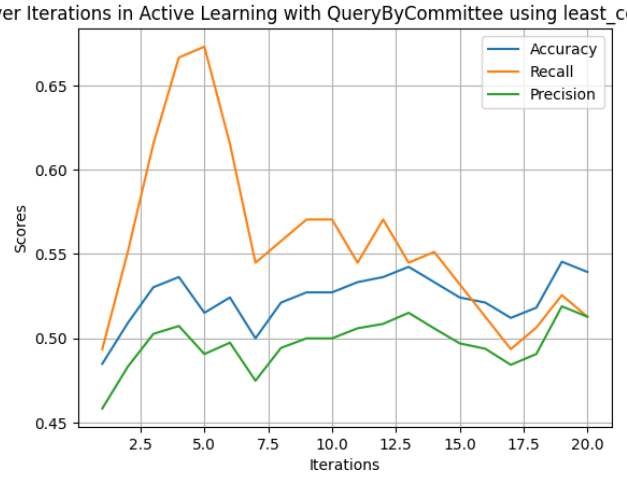
Traditional Ml : KNN Achieved accuracy of 50.60% , recall 47.77 % and precision 48.07%

Uncertainty Sampling: Achieved accuracy of 52.2% after 20 iterations.

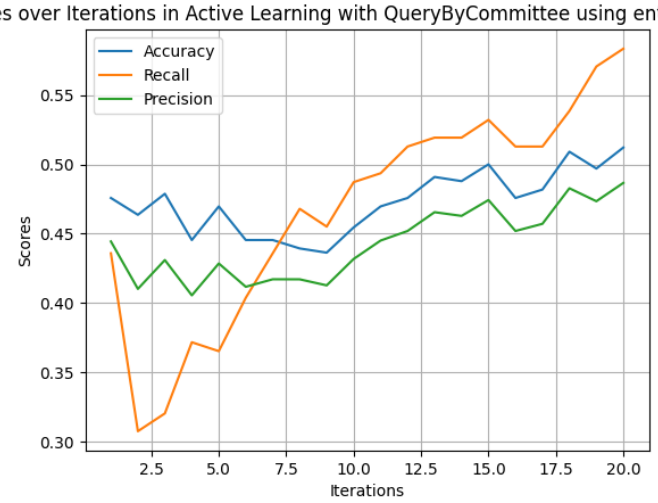
Entropy Sampling: Achieved accuracy of 51.2% after 20 iterations.

Margin Sampling: Achieved accuracy of 51.5% after 20 iterations.

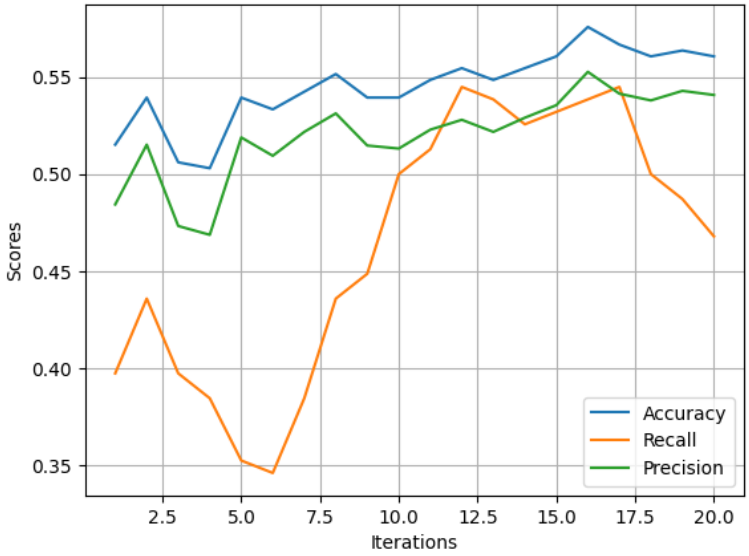
Scores over Iterations in Active Learning with QueryByCommittee using least_confident method



Scores over Iterations in Active Learning with QueryByCommittee using entropy method



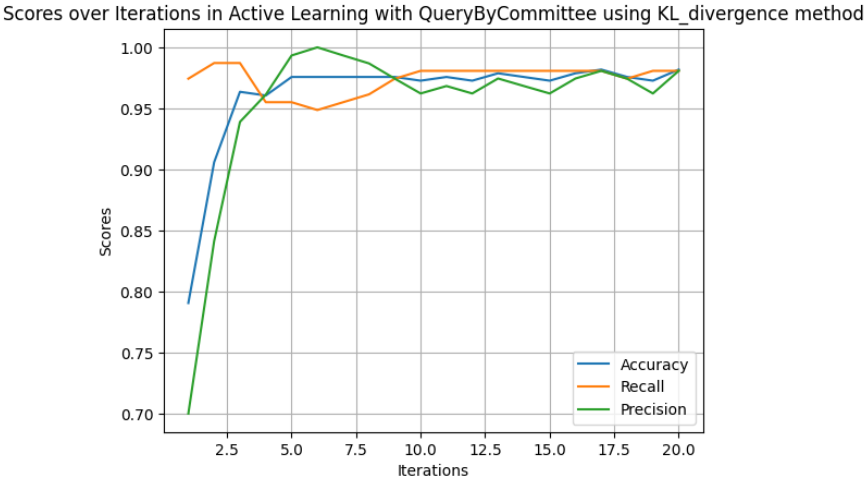
Scores over Iterations in Active Learning with QueryByCommittee using margin_sampling method

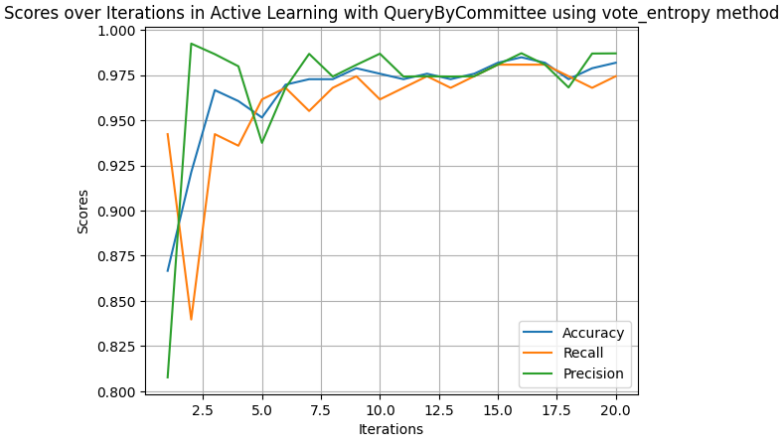


By using query by committee using (Vote entropy and KL divergent)

KL divergence : Achieved accuracy of 98.1 % after 20 iterations.

Vote entropy: Achieved accuracy of 97 % after 20 iterations.





The difference in accuracy between the traditional uncertainty sampling methods and query by committee methods can be attributed to the latter's ability to leverage model disagreement and capture more informative instances for learning. Query by committee methods, especially those using measures like KL divergence, are more effective at exploiting the diversity of the models in the committee, leading to substantial improvements in accuracy over iterations.

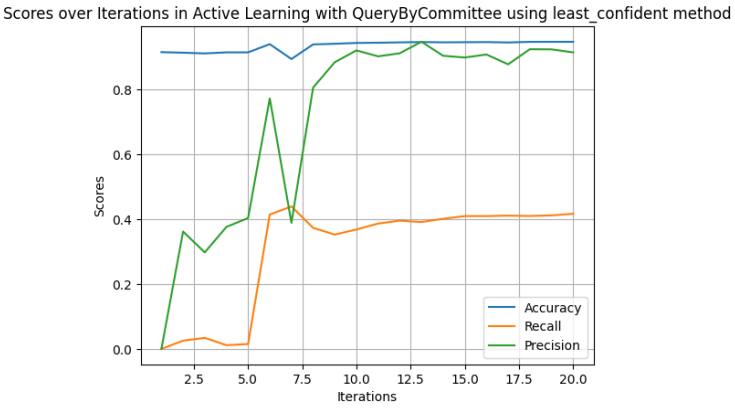
6.2. Diabetes Prediction Dataset Experiment

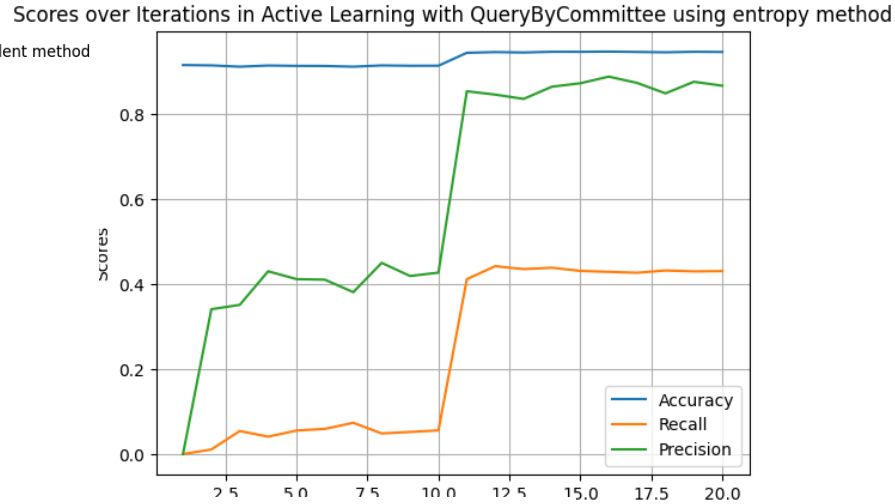
Traditional Ml : Random forest Achieved accuracy of 97.04 % , recall 95.25 % and precision 68.42%

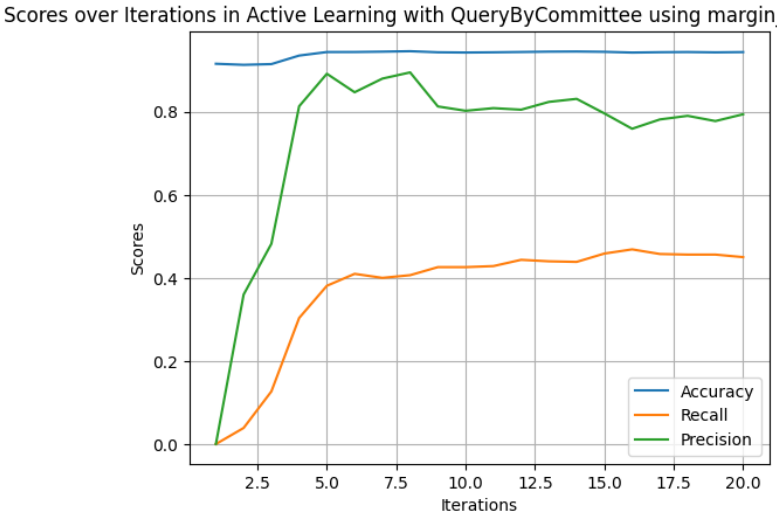
Uncertainty Sampling: Achieved accuracy of 94.8% after 20 iterations.

Entropy Sampling: Achieved accuracy of 94.6% after 20 iterations.

Margin Sampling: Achieved accuracy of 94.7% after 20 iterations.



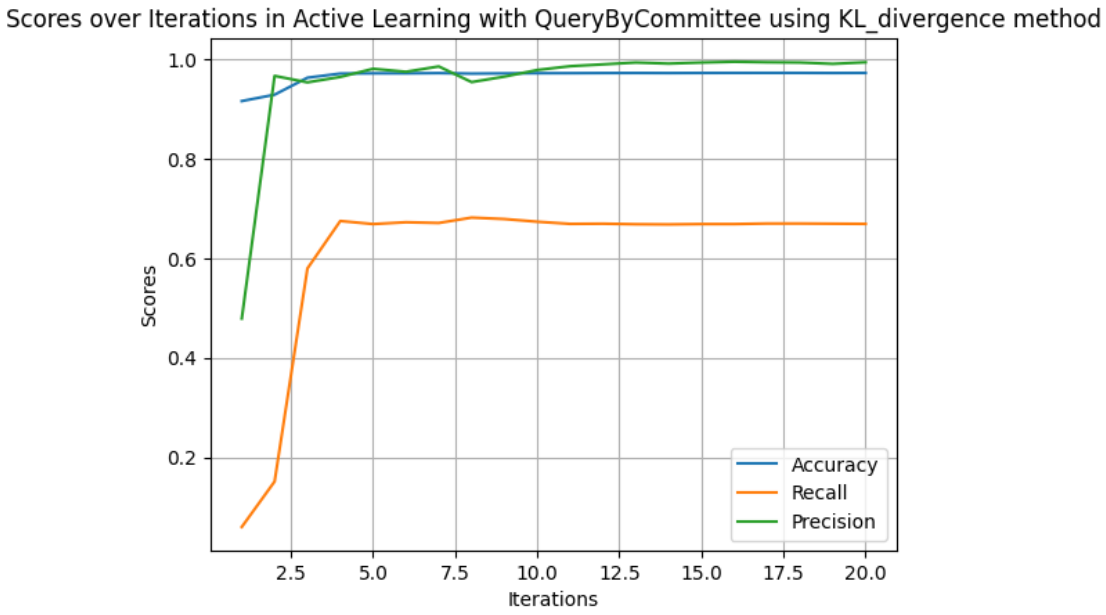


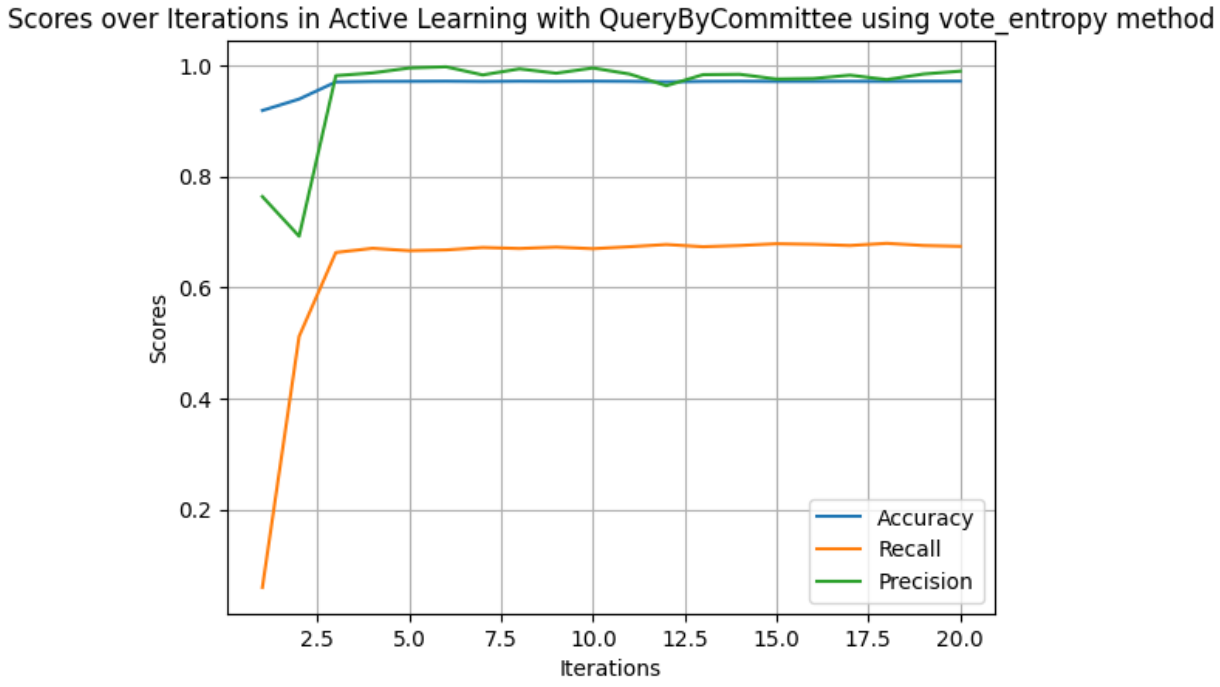


- ***By using query by committee using (Vote entropy and KL divergent)***

KL divergence : Achieved accuracy of 97.2 % after 20 iterations.

Vote entropy: Achieved accuracy of 97. 1% after 20 iterations.





The difference in accuracy between the traditional uncertainty sampling methods and query by committee methods can be attributed to the latter's ability to leverage model disagreement and capture more informative instances for learning. Query by committee methods, especially those using measures like KL divergence, are more effective at exploiting the diversity of the models in the committee, leading to substantial improvements in accuracy over iterations.

7. Conclusion

The experiments demonstrate the effectiveness of active learning strategies in improving model performance with limited labeled data. By iteratively selecting the most informative instances for labeling, we observe significant improvements in prediction accuracy across both the Loan Approval and Diabetes Prediction datasets. These findings underscore the potential of active learning techniques in enhancing machine learning models, particularly in domains where labeled data is scarce or costly to obtain.