



# Active Learning

*By*

*Abdelrahman Mostafa (20200827)*

*Abdelrahman Amin (20200311)*

*Mohamed Hisham (20200483)*

*Yousef Mohamed (20200669)*

Artificial Intelligence Department  
Cairo University  
2023-2024

## Table of Contents

<b>1. INTRODUCTION .....</b>	<b>5</b>
<b>2. OBJECTIVES .....</b>	<b>6</b>
<b>3. METHODOLOGY .....</b>	<b>7</b>
3.1. DATA.....	7
3.1.1. <i>Balanced Datasets</i> .....	7
3.1.1.1. <i>Iris</i> .....	7
3.1.1.2. <i>Digits</i> .....	7
3.1.2. <i>Imbalanced Datasets</i> .....	8
3.1.2.1. <i>BMI</i> .....	8
3.2. <i>Used Active Learning Scenarios</i> .....	8
3.2.1. <i>Pool-Based Active Learning</i> .....	8
3.3. <i>Used Machine Learning</i> .....	9
3.3.1. <i>Random Forest Model</i> .....	9
<b>4. EXPERIMENTAL RESULTS .....</b>	<b>11</b>
<b>4.1. BALANCED DATASET .....</b>	<b>11</b>
4.1.1. <i>BEFORE ACTIVE LEARNING</i> .....	11
4.1.2. <i>AFTER ACTIVE LEARNING</i> .....	11
<b>4.2. UNBALANCED DATASET .....</b>	<b>13</b>
4.2.1. .... <b>BEFORE BALANCING</b>	<b>13</b>
4.2.2. .... <b>AFTER BALANCING</b>	<b>13</b>
<b>5. CONCLUSION .....</b>	<b>14</b>
<b>5.1. BALANCED DATASET .....</b>	<b>14</b>
5.1.1. .... <b>IRIS</b>	<b>14</b>
5.1.2. .... <b>DIGITS</b>	<b>15</b>
<b>5.2. UNBALANCED DATASET .....</b>	<b>16</b>

5.2.1. ....	BMI	
.....		16
5.2.1.1. ....	ACTIVE LEARNING BEFORE BALANCING	
.....		16
5.2.1.2. ....	ACTIVE LEARNING AFTER BALANCING	
.....		17
6. GUI .....		18

# Abstract

Active learning is a powerful technique within machine learning (ML) that allows algorithms to achieve high accuracy with a minimal amount of labeled data. This report explores the core concepts of active learning, its advantages and disadvantages, and how it can be applied to various machine learning tasks.

Traditional supervised learning relies on passively consuming pre-labeled data. Active learning disrupts this paradigm, introducing a strategic approach that empowers the model to actively participate in its education. This abstract explores the core concepts of active learning in machine learning, highlighting its key benefits.

The core idea revolves around a pool-based learning strategy. The model starts with a limited set of labeled data and iteratively queries a vast pool of unlabeled data for the most informative points. These points are chosen based on a carefully selected query strategy, such as uncertainty sampling or committee disagreement. Human experts then label the chosen data points, significantly impacting the model's learning process. This creates an active learning loop: select, label, retrain, and repeat until the desired performance is achieved.

Active learning offers several compelling advantages. It can dramatically reduce the cost of labeling data, a significant bottleneck in many machine-learning projects. Focusing on informative data can also lead to improved model performance compared to a random sampling of the entire data pool. Additionally, active learning effectively leverages the power of unlabeled data, a valuable resource often underutilized in traditional approaches.

# Core Concepts

**Learner:** The machine learning algorithm responsible for selecting the most informative data points.

**Unlabeled Data Pool:** A vast collection of data points without corresponding labels.

**Oracle (Annotator):** A human expert or another information source that provides labels for the selected data points.

Active learning boils down to strategically selecting the most informative data points for labeling from a large pool of unlabeled data. Here's a quick rundown of the key ideas:

- **Pool-based learning:** Train on a small set of labeled data (seed), then iteratively query the unlabeled pool for informative points based on a chosen strategy.
- **Query strategy:** This is the "how" of choosing informative points. Examples include uncertainty sampling (focusing on unsure predictions) and query by committee (exploiting disagreements among multiple models).
- **Human-in-the-loop (HITL):** Humans provide labels for the chosen data points, impacting model performance significantly.
- **Active learning loop:** This cycle defines the process: select -> label -> retrain -> repeat until a stopping point is reached.
- **Model uncertainty & data distribution:** These concepts influence query strategies. Uncertainty focuses on the model's confidence, while data distribution ensures a diverse training set.

# 1. Introduction

## The Power of Selective Learning: Introducing Active Learning in Machine Learning

Machine learning thrives on data, but not all data is created equal. Labeling data, especially for complex tasks, can be a time-consuming and expensive bottleneck. This is where Active Learning steps in, offering a smarter approach to training machine learning models.

Active learning flips the script on traditional supervised learning. Instead of passively consuming a pre-labeled dataset, the model actively participates in its education. It strategically selects the most informative data points from a large pool of unlabeled data, presenting them to a human expert for labeling. These carefully chosen examples help the model learn more effectively, often achieving better performance with significantly less labeled data.

This introduction sets the stage for exploring the exciting world of Active Learning. We'll delve deeper into the specific objectives of using this approach, such as reducing labeling costs, improving model performance, and leveraging the power of unlabeled data. By strategically choosing what to learn from, Active Learning empowers us to build more efficient and powerful machine-learning models.

## 2.Objectives

There are several key objectives for applying active learning in machine learning:

### 1. Reduce Labeling Cost:

- Labeling data can be a significant bottleneck in machine learning projects, often requiring human expertise and time. Active learning aims to strategically select the most informative data points for labeling, significantly reducing the amount of human effort required to achieve good model performance. This is especially valuable when dealing with large datasets where labeling everything is impractical.

### 2. Improve Model Performance:

- By focusing on informative data points, active learning allows the model to learn from the most valuable examples. These points often lie in areas of uncertainty for the model or represent critical boundaries between classes. This targeted learning approach can lead to better model performance compared to training on a random sample of the entire data pool.

### 3. Leverage Unlabeled Data:

- Real-world data often comes with a large amount of unlabeled data and a smaller set of labeled data. Active learning allows us to effectively utilize this unlabeled data by strategically incorporating it into the training process. This can be particularly advantageous when labeled data is scarce or expensive to obtain.

### 4. Reduce Training Time:

- Since active learning requires labeling fewer data points, it can potentially lead to faster training times for machine learning models. This is especially true for complex models that require a significant amount of data to train effectively.

### 5. Targeted Data Collection:

- Active learning can be used to guide data collection efforts. By identifying the most informative data points, active learning can help focus data collection on areas that will have the biggest impact on model improvement. This is useful in scenarios where actively gathering new data is possible.

In summary, active learning offers a strategic approach to training machine learning models by focusing resources on the most valuable data points. This leads to reduced labeling costs, improved model performance, and efficient utilization of data.

## 3. Methodology

### 3.1. Data

#### 3.1.1. Balanced Datasets

##### 3.1.1.1. Iris

The Iris flower dataset is a classic dataset used in machine learning for tasks like classification. It consists of measurements of 150 iris flowers from three different species: Iris setosa, Iris versicolor, and Iris virginica.

Here's a breakdown of the dataset:

- Number of samples: 150 (50 from each species)
- Features: 4 features are measured for each flower: sepal length, sepal width, petal length, and petal width (all in centimeters)
- Target: The target variable is the iris species (Setosa, Versicolor, or Virginica)

This dataset is valuable because:

- It's a small and well-understood dataset, making it a good starting point for learning machine learning algorithms.
- It involves a classification task, which is a common problem in machine learning.
- Despite its simplicity, the data can expose challenges like class separation, where some species are more easily distinguished than others.

##### 3.1.1.2. Digits

The scikit-learn handwritten digits dataset is a collection of images of handwritten digits (0-9). It's a popular dataset for beginners to practice image classification tasks. Here's a breakdown of its key features:

- Images: The dataset consists of 1797 grayscale images of handwritten digits, each 8x8 pixels in size.
- Features: Each image is essentially a 2D array representing the pixel intensities. For machine learning algorithms, these images need to be flattened into a 1D feature vector of length 64 (8 pixels x 8 pixels).
- Target: The dataset includes a target variable that labels each image with the corresponding digit (0-9).



This dataset is useful for:

- Learning image classification: By training a model on this dataset, you can explore how algorithms classify images based on their pixel features.
- Simple and interpretable: The dataset is relatively small and the images are easy to visualize, making it a good choice for understanding image classification concepts.
- Benchmarking: The dataset is a common benchmark for evaluating the performance of image classification algorithms.

### 3.1.2. Imbalanced Datasets

#### 3.1.2.1. BMI

BMI is based on Gender, Height & Weight.

The complexity arises because the dataset has fewer samples and is highly imbalanced.

This data frame contains the following columns:

- Gender: Male / Female
- Height: Number (cm)
- Weight: Number (Kg)
- Index :
  - Extremely Weak
  - Weak
  - Normal
  - Overweight
  - Obesity
  - Extreme Obesity

## 3.2. Used Active Learning Scenarios

### 3.2.1. Pool-Based Active Learning

Pool-based active learning is a strategy within the field of active learning used to efficiently train a machine learning model by strategically choosing which data points to label. Here's a breakdown of its key aspects:

Core Idea:

- Starts with a large pool of unlabeled data.
- The model iteratively selects the most informative data points from the pool for human annotation (labeling).
- These labeled points are used to retrain the model, which then guides the selection of the next informative points.
- This cycle continues until the desired model performance is achieved, requiring less labeled data compared to training on a random sample of the entire pool.

Benefits:

- **Reduced Labeling Cost:** By strategically choosing informative data points, pool-based active learning can significantly reduce the amount of human effort required for labeling data, which can be expensive and time-consuming.
- **Improved Model Performance:** Focusing on informative data points allows the model to learn more effectively, potentially leading to better performance compared to training on randomly chosen data.

Types of Pool-Based Sampling Strategies:

There are various approaches to selecting informative data points from the unlabeled pool. Some common strategies include:

- **Uncertainty Sampling:** The model prioritizes data points for which it has the least confidence in its prediction. These points are most likely to help the model learn and improve its accuracy.
- **Query by Committee (QBC):** The model trains multiple copies of itself (committee) and selects data points where the committee disagrees the most. This disagreement suggests a challenging case that can improve the committee's overall decision-making.
- **Density-weighted sampling:** This strategy focuses on selecting data points from areas of the feature space where there is a higher concentration of unlabeled data. This helps ensure the model is trained on a diverse set of examples.

Overall, pool-based active learning is a valuable technique for maximizing model performance while minimizing the need for labeled data.

### 3.3. Used Machine Learning

#### 3.3.1. Random Forest Model

A Random Forest model is a powerful and versatile machine learning algorithm that excels at both classification (predicting categories) and regression (predicting continuous values). It works by combining the predictions of multiple decision trees, creating an ensemble learner that is typically more robust and accurate than any single decision tree.

Here's a breakdown of how Random Forests work:

- **Building Decision Trees:** The algorithm generates a large collection of decision trees. Each tree is trained on a random subset of the training data, with some features randomly chosen at each split point within the tree. This randomness helps prevent the trees from becoming overly specialized on the training data (overfitting).
- **Voting for Classification (or Averaging for Regression):** For classification tasks, when a new data point arrives, it's passed through all the trees in the forest. Each tree votes for the class it thinks the data point belongs to. The final prediction is the class that receives the most votes from the individual trees. In regression tasks, the trees predict a continuous value, and the final prediction is the average of the predictions from all the trees.

Benefits of Random Forests:

- **High Accuracy:** By combining multiple decision trees, Random Forests can often achieve higher accuracy than single decision trees, especially on complex problems.
- **Robust to Overfitting:** The random injection of features at each split point in the trees helps prevent the model from memorizing the training data and reduces the likelihood of overfitting.
- **Handles Missing Data:** Random Forests can inherently deal with missing data in the training set, making them less prone to errors when encountering such data.
- **Can handle both Categorical and Continuous Features:** Random Forests can work effectively with datasets containing both categorical (nominal or ordinal) and continuous features, providing flexibility in data types.

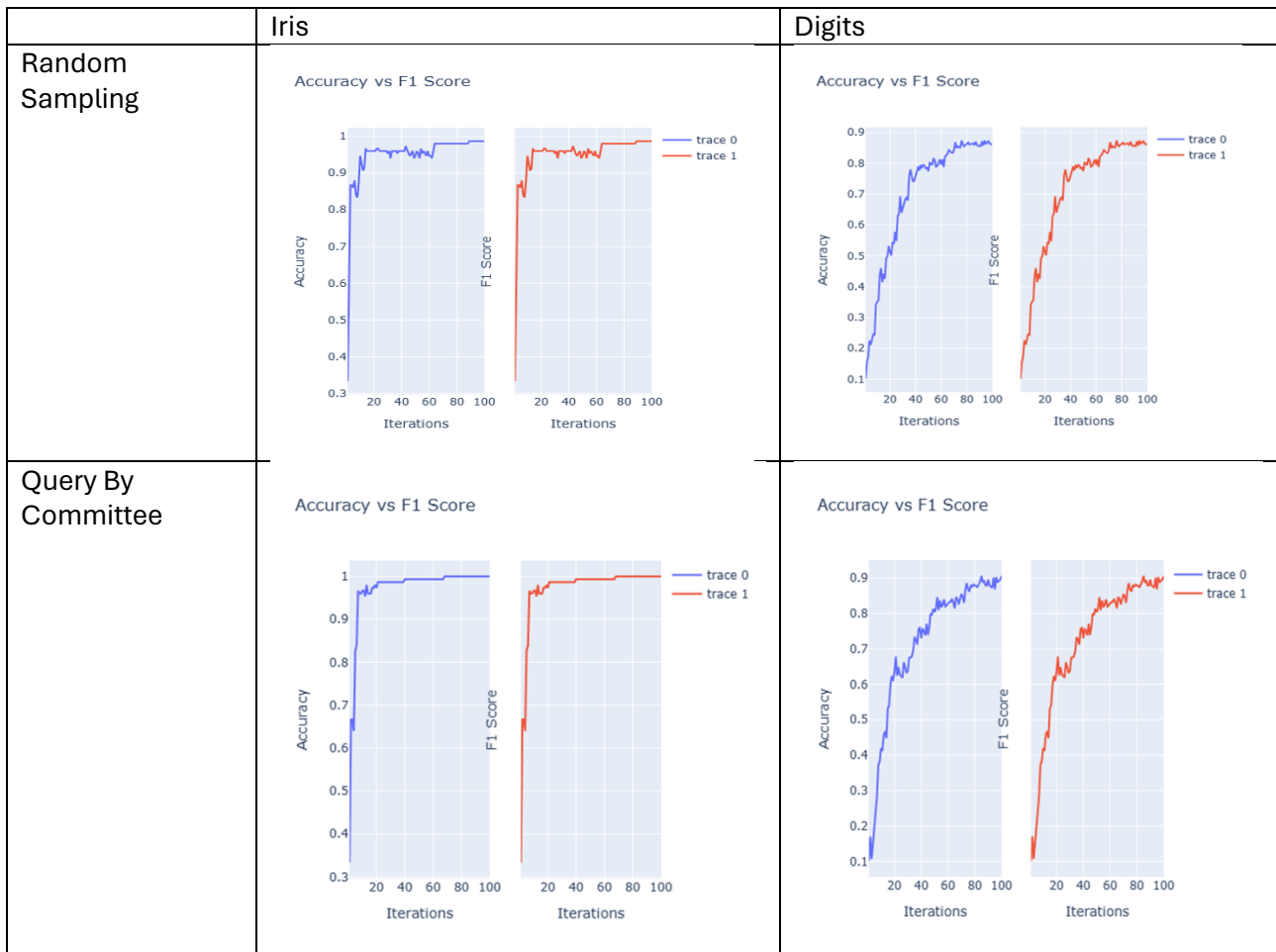
## 4. Experimental Results

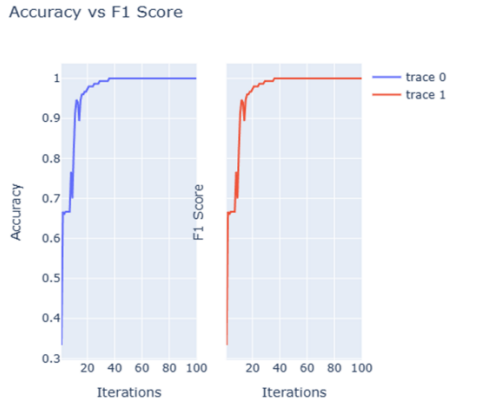
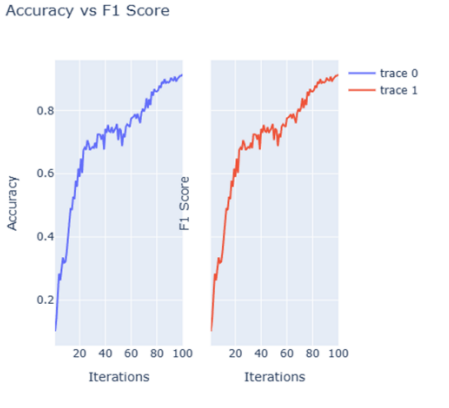
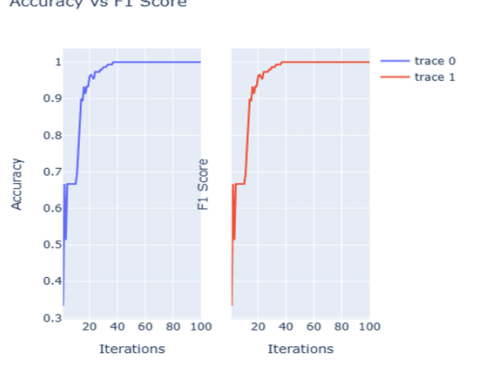
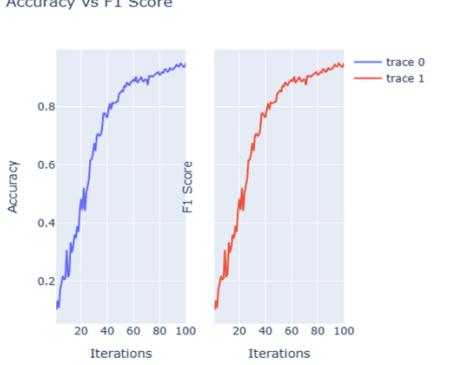
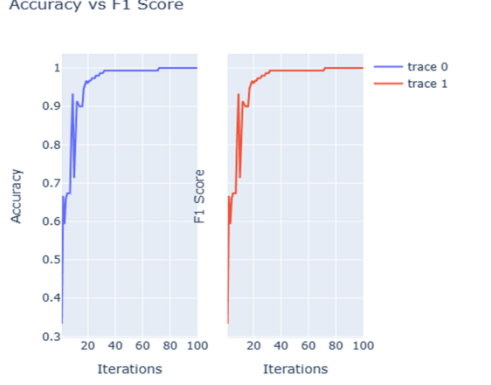

### 4.1. Balanced Dataset

#### 4.1.1. Before Active Learning

	Iris	Digits
Accuracy	98%	93%
F1 Score	~98%	~93%

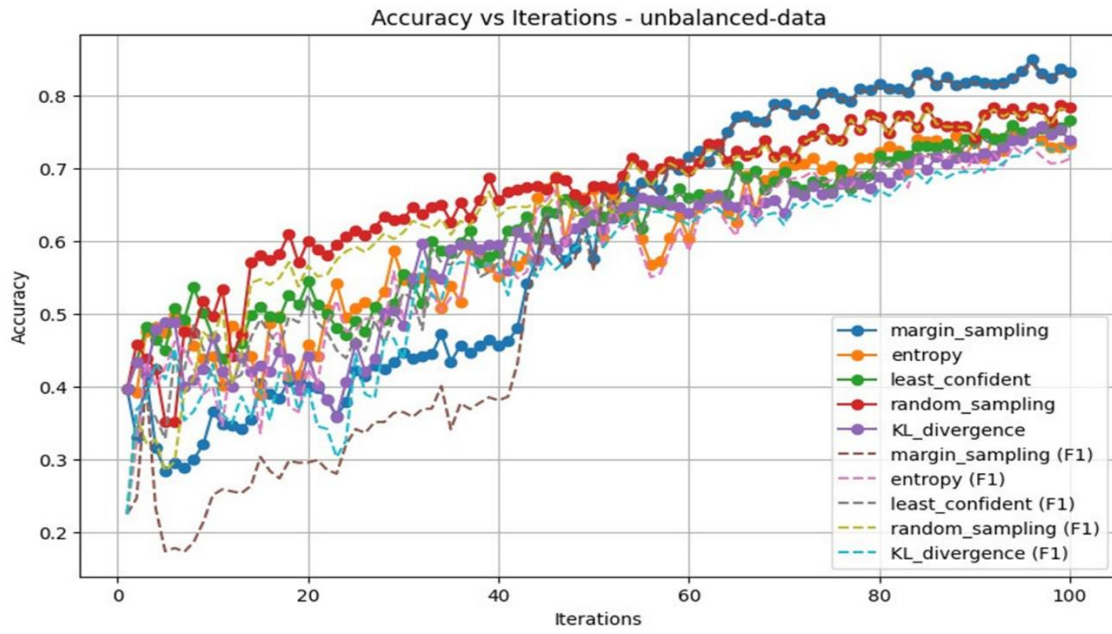
#### 4.1.2. After Active Learning



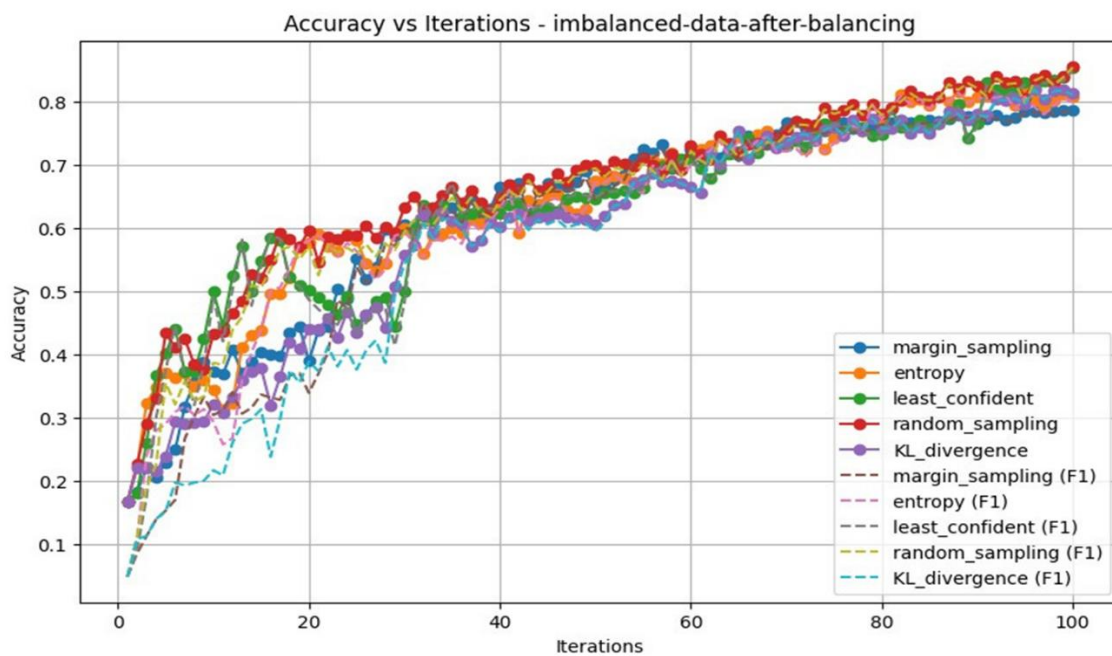
<p>Uncertainty Sampling with Least Confident</p>	<p>Accuracy vs F1 Score</p> 	<p>Accuracy vs F1 Score</p> 
<p>Uncertainty Sampling with Margin Sampling</p>	<p>Accuracy vs F1 Score</p> 	<p>Accuracy vs F1 Score</p> 
<p>Uncertainty Sampling with Entropy</p>	<p>Accuracy vs F1 Score</p> 	<p>Accuracy vs F1 Score</p> 

## 4.2. Unbalanced Dataset

### 4.2.1. Before Balancing



### 4.2.2. After Balancing

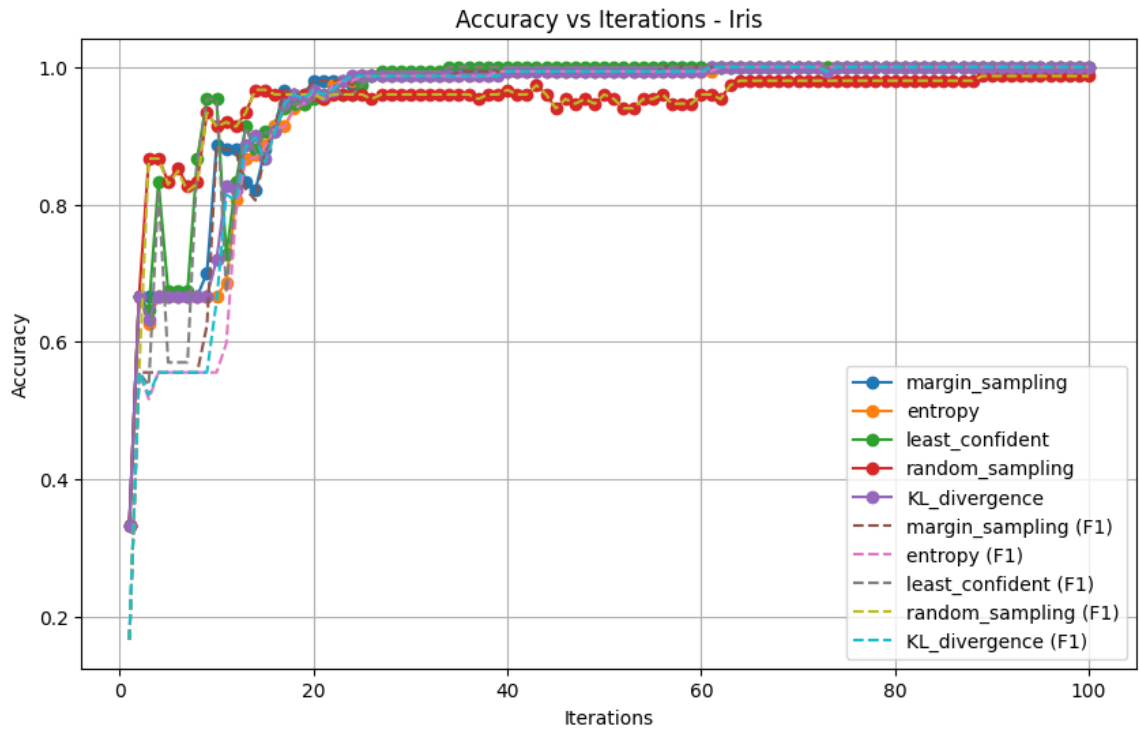


## 5. Conclusion

### 5.1. Balanced Dataset

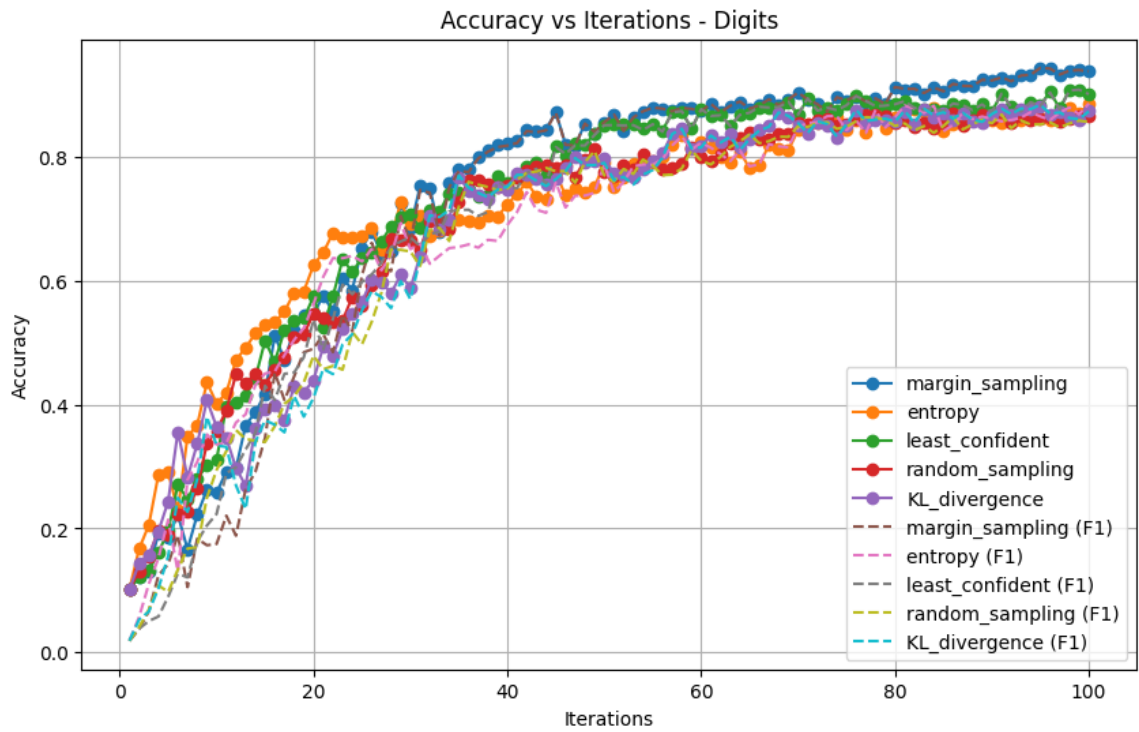
#### 5.1.1. Iris

	F1 Score	Accuracy
Random Sampling	~98%	98%
Query By Committee	~100%	100%
Uncertainty Sampling with Least Confident	~100%	100%
Uncertainty Sampling with Margin Sampling	~100%	100%
Uncertainty Sampling with Entropy	~100%	100%



## 5.1.2. Digits

	F1 Score	Accuracy
Random Sampling	85.66%	86.53%
Query By Committee	87.38%	87.42%
Uncertainty Sampling with Least Confident	90.17%	90.15%
Uncertainty Sampling with Margin Sampling	93.86%	93.82%
Uncertainty Sampling with Entropy	88.66%	88.59%



The F1 score and accuracy might be the same or very similar in all iterations if the dataset is balanced and the model is performing well. The Iris dataset is a balanced dataset and simple dataset, meaning it has an equal number of samples for each class. In such cases, accuracy and F1 score can be very similar. The Digits dataset is a balanced dataset but more complex than the Iris dataset. So, accuracy and F1 score have different values unlike in Iris Dataset.

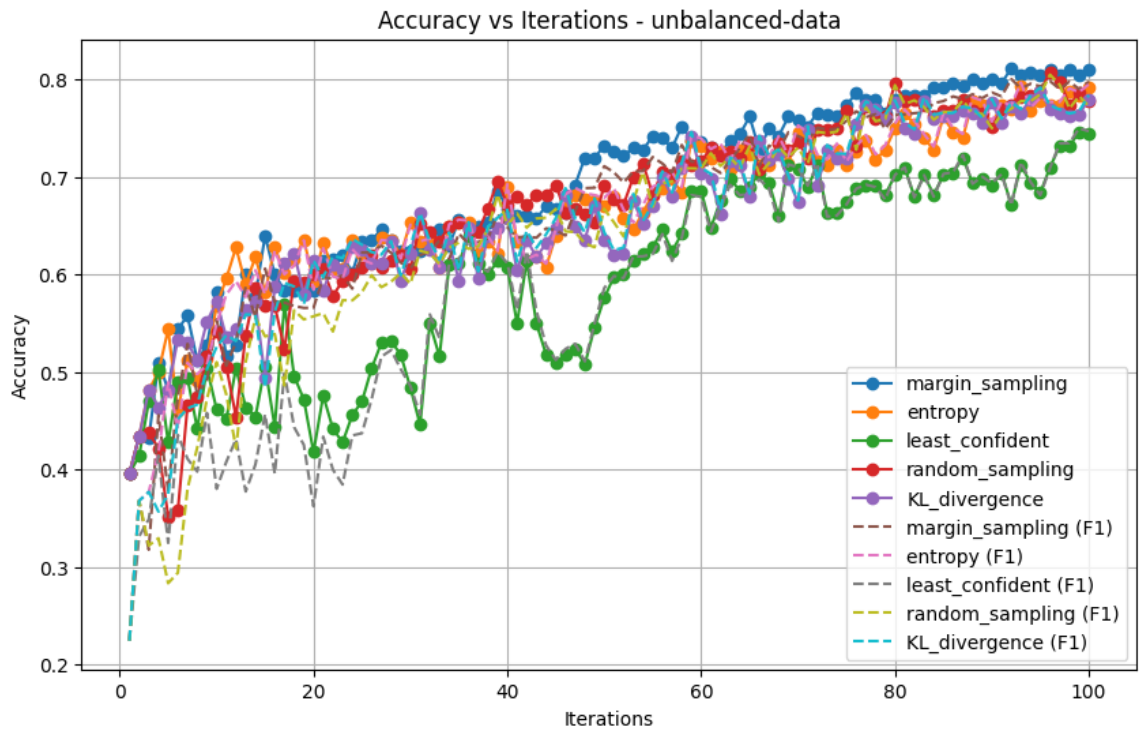


## 5.2. Unbalanced Dataset

### 5.2.1. BMI

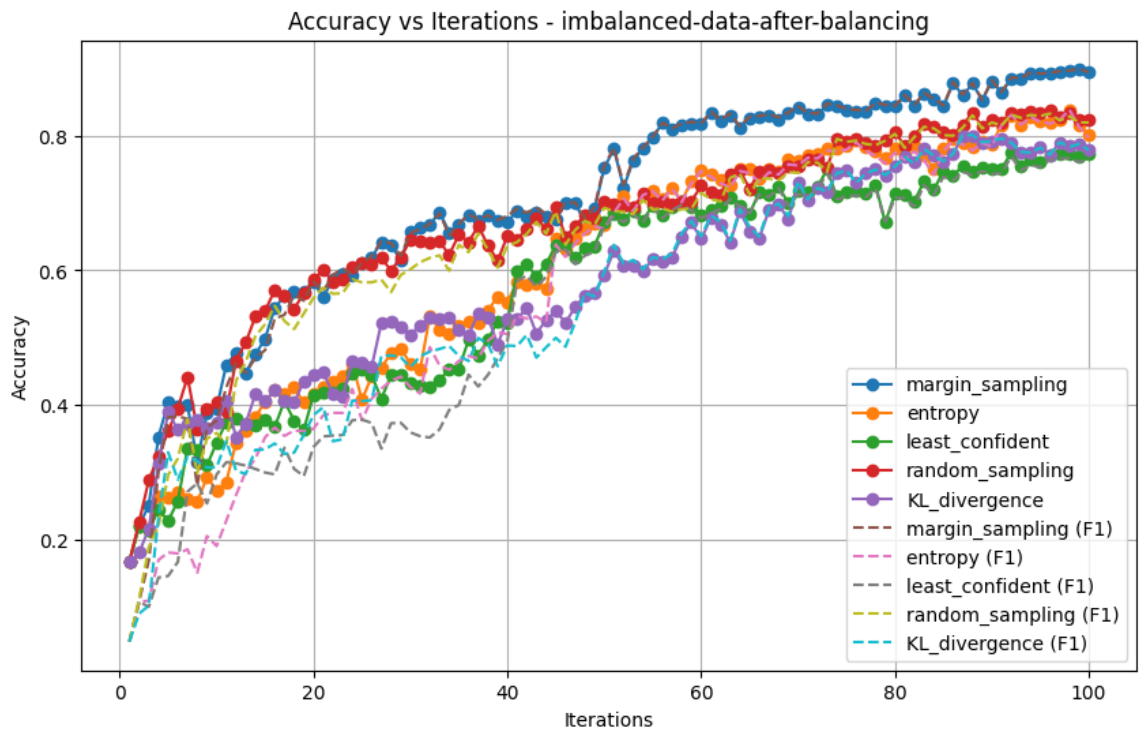
#### 5.2.1.1. Active Learning Before Balancing

	F1 Score	Accuracy
Random Sampling	77.28%	77.8%
Query By Committee	78.39%	78%
Uncertainty Sampling with Least Confident	74.7%	74.4%
Uncertainty Sampling with Margin Sampling	79.79%	81%
Uncertainty Sampling with Entropy	79.51%	79.2%

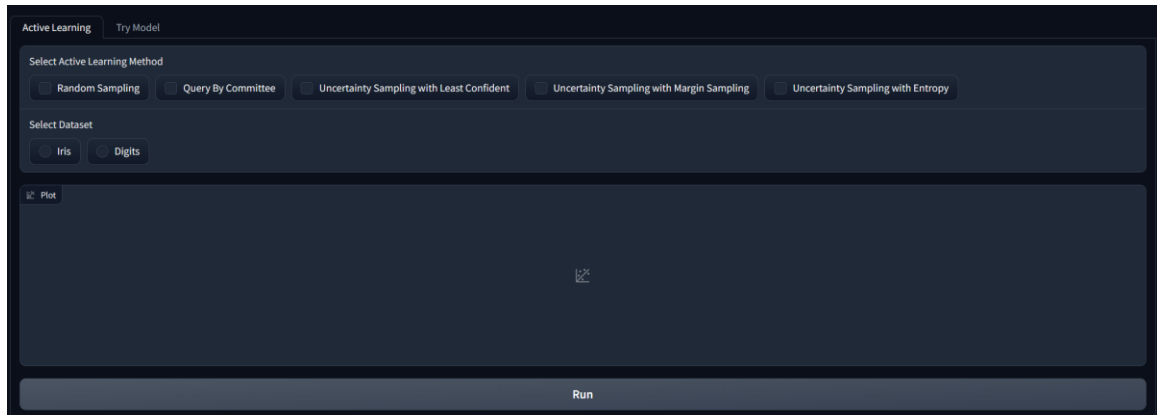


### 5.2.1.2. Active Learning After Balancing

	F1 Score	Accuracy
Random Sampling	81.97%	82.32%
Query By Committee	77.89%	77.86%
Uncertainty Sampling with Least Confident	76.68%	77.36%
Uncertainty Sampling with Margin Sampling	89.47%	89.39%
Uncertainty Sampling with Entropy	79.72%	80.13%



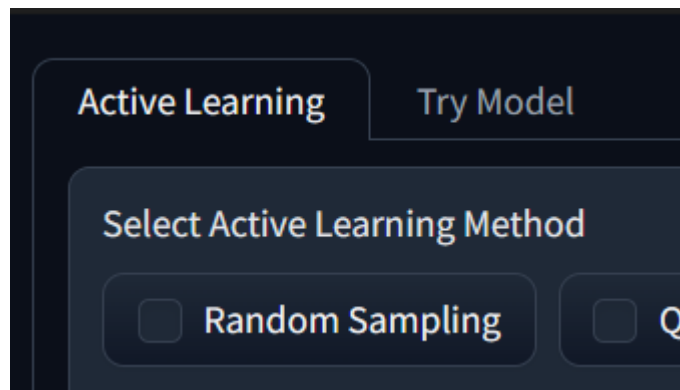
## 6. GUI



You have two pages in our GUI.

The first page is called 'Active Learning' which is the place where you can use different Active Learning methods.

In the 'Active Learning Method' section, you can select any number of methods provided in this section at the same time producing a beautiful graph of accuracy and f1 score.



The second page is called 'Try Model' which is the place where you can try the model that you applied Active Learning methods on it.

You must apply any Active Learning method to try the model.