# Task 1

## Data preparation

Before we can start solving the tasks we need to prepare the input data. As in mini project 2, we use the answerList_data.csv file. We create a column `correctness` that specifies if the ground truth and the answer of the participant match. We also need to transform some of the columns to values, we can use in our classifier later and calculate the TTR (token-type-ratio) of each explanation. Then we split the data into a training and a holdout set and those again into student and non-student data. We balance the training data to have the same number of correct and incorrect answers. We also only use the input columns `Worker.score`, `Answer.duration`, `TTR`, `explanation_size`, `Answer.confidence`, and `Answer.difficulty` since they were given in the task.

Out[2]:

| | Answer.ID | FailingMethod | Question.ID | Answer.duration | Answer.confidence |
|---|---|---|---|---|---|
| **0** | 261 | HIT01_8 | 0 | 90.984 | 4 |
| **1** | 262 | HIT01_8 | 0 | 133.711 | 5 |
| **2** | 263 | HIT01_8 | 0 | 77.696 | 5 |
| **3** | 264 | HIT01_8 | 0 | 46.644 | 1 |
| **4** | 265 | HIT01_8 | 0 | 215.416 | 5 |
| **...** | ... | ... | ... | ... | ... |
| **2575** | 2316 | HIT08_54 | 128 | 220.420 | 2 |
| **2576** | 2317 | HIT08_54 | 128 | 322.790 | 4 |
| **2577** | 2318 | HIT08_54 | 128 | 159.530 | 5 |
| **2578** | 2319 | HIT08_54 | 128 | 68.578 | 5 |
| **2579** | 2320 | HIT08_54 | 128 | 72.605 | 4 |

2580 rows × 26 columns

Out[7]:
```
correctness
0    297
1    297
dtype: int64
```

Out[8]:
```
correctness
0    562
1    562
dtype: int64
```

Out[9]:
```
correctness
0    34
1    34
dtype: int64
```

Out[10]:
```
correctness
0    62
1    62
dtype: int64
```

---

# Evaluating the classifier on student and non-student data

With our data prepared we can train a RandomForestClassifier on the student-only train data set. The following results show the precision, recall, and F1 score for the student-only holdout set and the non-student holdout set.

```
Student precision: 0.574468085106383
Student recall: 0.7941176470588235
Student F1 score: 0.6666666666666666
Non-student precision: 0.5232558139534884
Non-student recall: 0.7258064516129032
Non-student F1 score: 0.6081081081081081
```

Here the results are as expected, the model performs better on the student data than on the non-student data.

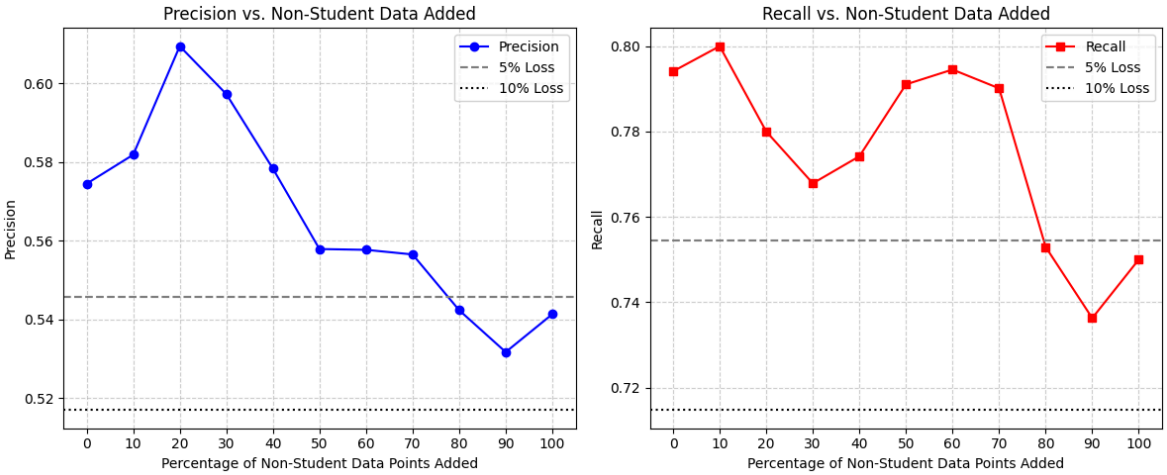
---

# Question 1: Gradually adding more non-students to the holdout set

To answer question 1, we gradually add non-student data from the holdout set to the holdout set with student data. For every 10% added we generate new predictions and calculate the precision and recall. We plot the results to see if there is a clear trend.

We see a downward trend for the precision if we gradually add more and more non-student data to the holdout set. For the precision, a loss of 5% is reached with 80% non-student data added. With more non-student data, it further decreases, but does not reach a 10% loss.

For the recall we do not see such a clear trend as it varies more. After adding 80% of the non-student data, we briefly reach a 5% loss in recall. As you can also see in the plot, the precision never reaches a 10% loss.




---

# Question 2: Gradually adding non-students to train data

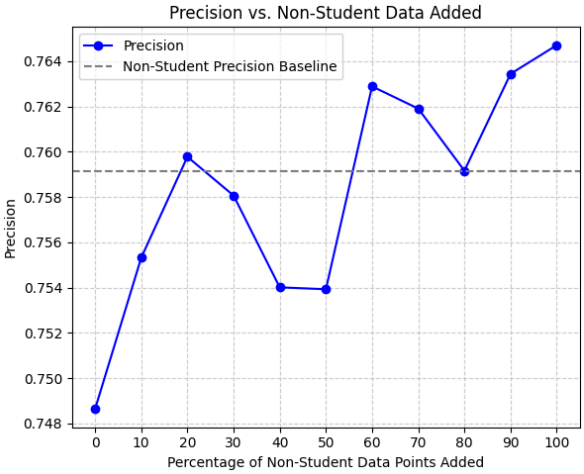
Now we do not add non-student data to the holdout set, but we gradually add it to the training. This way we want to find out, when we reached the precision and recall scores of the original model when using non-student data as the holdout set.

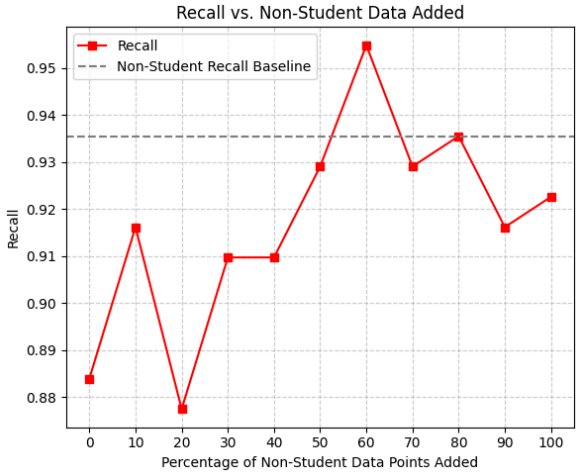
Before doing that we need to train a model on all the training data as a reference. Therefor we use the RandomForestClassifier model here and train it on all train data. This way we can get a baseline that is comparable to the results of mini project 2. The predictions of the non-student holdout set on this trained model will then be our baseline for the comparison. Those baselines will be displayed in the plots as dashed lines.

Now we can gradually add non-student data to the training data and evaluate the model on the holdout set. We plot the results to see if there is a clear trend.

Here again, we can see an upward trend. Precision and recall increase, as we add more non-student data to the training set. Although this trend is very volatile, we can clearly see a tendency here. But we can see from these plots, that when adding 20% of the non-student data to the training set, we already reach similar precision and recall scores of the original model. This is interesting because we only have 20% of the non-student data in the training set, but we already reach the same performance as the original model. For the recall the story is a bit different. Only from 50% to 60% non-student data added, we reach a similar recall score as the original model.

This in return means that we need to add about a half of the non-student data to the training set to reach the same performance as the original model.


Precision vs. Non-Student Data Added
Precision
Non-Student Precision Baseline
Precision
Percentage of Non-Student Data Points Added
0.764
0.762
0.760
0.758
0.756
0.754
0.752
0.750
0.748
0
10
20
30
40
50
60
70
80
90
100


Recall vs. Non-Student Data Added
Recall
Non-Student Recall Baseline
Recall
Percentage of Non-Student Data Points Added
0.95
0.94
0.93
0.92
0.91
0.90
0.89
0.88
0
10
20
30
40
50
60
70
80
90
100