Lab 1 – Theoretical Tasks

Task 1.1 Ethics

Given professions translated from English (neutral) into German (non-neutral) with ChatGPT.

- The doctor – der Arzt
- The nurse – die Krankenschwester
- The lawyer – der Anwalt
- The office worker – der Büroangestellte
- The janitor – der Hausmeister
- The construction worker – der Bauarbeiter

ChatGPT matched the grammatical gender (der – male/die - female) to the stereotypical gender often associated with each profession. This does not mean that f.e. there are just male doctors (der Arzt), but doctors are still often imagined as male in old-school stereotypes.

Task 1.2 Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Example 1:

- True Negatives (TN): 990
- False Positives (FP): 10
- False Negatives (FN): 20
- True Positives (TP): 30

		Predicted	
		Negative	Positive
Actual	Negative	990	10
	Positive	20	30

Calculate accuracy, precision, recall, and F1-score:

Accuracy = (30 + 990) / (990 + 10 + 20 + 30) = 0,97

Precision = 30 / (30 + 10) = 0,75

Recall = 30 / (30 + 20) = 0,6

F1-score = 2 * 0,75 * 0,6 / (0,75 * 0,6) = 0,67

Accuracy is not a good measurement in that case, because the 20 FN is quite significant compared to the 30 TP which shows that the classifier struggles with correctly classifying the positive class but since the positive class is much smaller than the negative class the positive class, especially the FN do not carry any weight. This observation shows that accuracy as a metric is not useful for imbalanced data. Precision, recall and F1-score are a much better representation of the classifier's performance.

Example 2:

- TN: 9000
- FP: 50
- FN: 100
- TP: 850

		Predicted	
		No	Yes
Actual	No	9000	50
	Yes	100	850

Calculate accuracy, precision, recall, and F1-score:

Accuracy = (850+ 9000) / (9000 + 50 + 100 + 850) = 0,985

Precision = 850 / (850 + 50) = 0,94

Recall = 850 / (850 + 100) = 0,895

F1-score = 2 * 0,94 * 0,895 / (0,94 * 0,895) = 0,92

Even though the classes are still unbalanced the number of falsely predicted labels compared to the correctly predicted ones is fairly good. But it seems that predicting the Yes category correctly is still harder than the No category. This is why the Accuracy is still higher than the other metrices.

When Accuracy is not the best metric:

1. **Class Imbalance:**
   In both cases the negative class dominates (1000 vs 50 in the first case, 9050 vs 950 in the second). Even though Accuracy is high in both cases (97,1% and 98,5%), the other metrics like Precision, Recall, and F1-Score are much lower. The problem lies in the formula of accuracy: FNs do not carry any weight when we calculate Accuracy.
2. **Unequal Error Costs:**
   Accuracy aggregates all correct predictions without distinguishing between FP and FN. In a medical diagnosis scenario, FN (missing a disease) may be more critical than FP (a false alarm).
3. **Limited Insight into Model Performance:**
   Accuracy does not indicate where the model is making mistakes. In the second case is Recall 9%-points lower than Accuracy, meaning that actual positive ones are missed.