

# INTRODUCTION TO DATA SCIENCE WS25/26

---

## Report Assignment Part 1

---

*Group 85:*

Maurice Chartée 434945

Ewa Schönborn 422197

## **Statement on the usage of LLMs**

We used an LLM for grammar, spell checking and sentence structure in Tasks 2,4,5,6. The aforementioned usage of an LLM does not detract from our understanding of the tasks, since all analytical decisions, the selection and application of tools, and interpretation of results were carried out independently before involving the LLM.

## Q1 Data Exploration

- a) The data frame has 7 columns and 1349 rows.

	age	bmi	children	charges
count	1348.00	1347.00	1348.00	1347.00
mean	39.23	30.66	1.10	13254.72
std	14.06	6.09	1.22	12096.11
min	18.00	15.96	0.00	1121.87
25%	27.00	26.32	0.00	4742.31
50%	39.00	30.36	1.00	9377.90
75%	51.00	34.64	2.00	16582.14
max	64.00	53.13	7.00	63770.43

Figure 1: Table showing the basic statistics resulting from the `describe()`-method

b) The basic statistics can be found in Figure 1. The `describe`-method by default only calculates basic statistics for columns with numerical values. The columns not contained in the table are non-numerical and therefore excluded.

c) The row count of the data frame without any `Nan` values is 1338. We continue with this data frame.

d) There are 20 underweight patients, 225 normal weight patients, 386 overweight patients and 707 obese patients.

As we do not learn from the data yet and predict the BMI, we do not balance out the data.

If we would balance out the data set, we could use stratified sampling or undersampling.

e) The histogram in Figure 2 visualizes the distribution of values of the `children` feature. Its mode is 0.

f) The histogram in Figure 3 shows the resulting histogram with 40 bins. The number was chosen to provide enough information to be informative while reducing the number of empty bins and bins with few instances. The histogram is an equal-width histogram with 40 bins that is multi-modal and right-skewed.

## Report Assignment Part 1

---

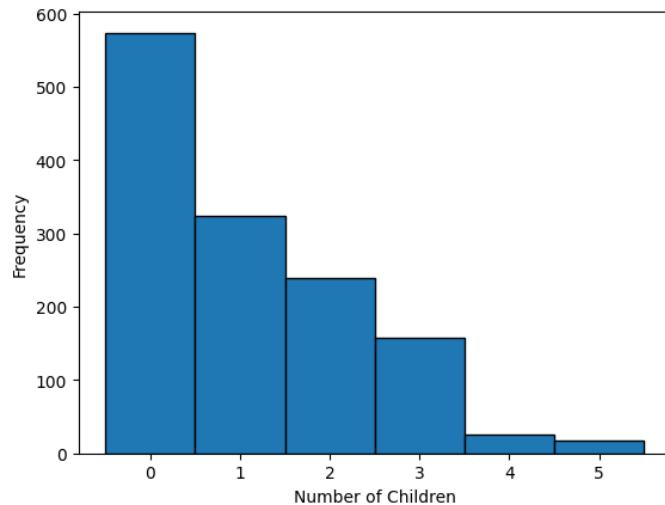


Figure 2: Histogram of the feature 'children'

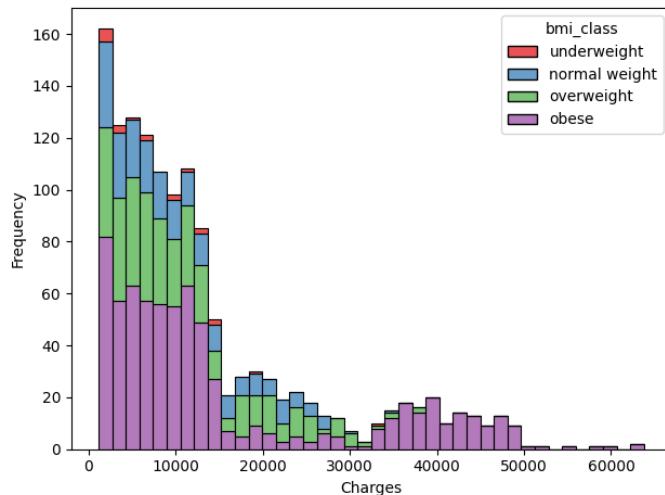


Figure 3: Equal-width histogram of the feature 'charges'

- g)** The higher the charges for the patients are, the less likely it is for the patient to be not obese. If the person is underweight, the charges tend to be lower.
- h)** The scatter plot can be seen in Figure 4. We observe from the orange dots being shifted to the right that along all features the charges for patients that are smoking are generally higher. Being obese on top of being a smoker additionally increases the charges.

## Report Assignment Part 1

---

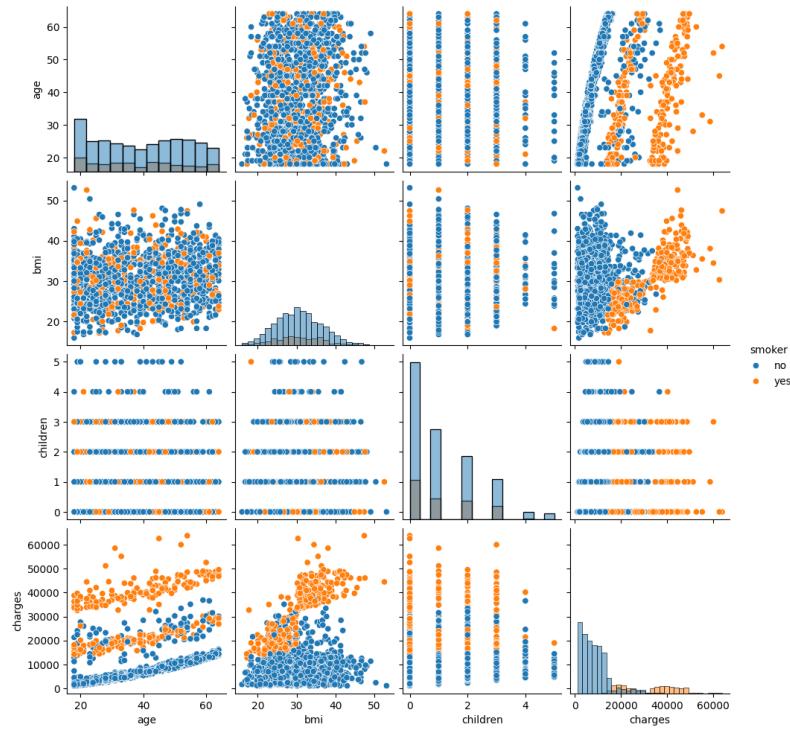


Figure 4: Scatter plot matrix for the numerical features of the dataset colored by smoke status

	age	bmi	children	charges
age	1.00	0.11	0.04	0.30
bmi	0.11	1.00	0.01	0.20
children	0.04	0.01	1.00	0.07
charges	0.30	0.20	0.07	1.00

Figure 5: Correlation matrix for the numerical features of the dataset

i) Figure 5 shows the correlation matrix. The strongest absolute correlation is 0.30 between age and charges. The slight positive correlation implies that the older a patient is, the higher the charges. However it should be considered that the correlation is not very strong, so age only slightly influences the charges.

j)

- Cannot be answered as the box plot only shows distributions and not single cases, so no statement about absolute frequencies can be made. However

## Report Assignment Part 1

---

proportionally to the size of the class, there are more smokers with charges above 20,000.

- No, as there appear no circles marking outliers on the left side of the lower fence
- Cannot be answered as the mean cannot be inferred from the box plot which shows only distributions and no single instances
- Yes, you can clearly see this in plot.

## Q2 Decision Tree

a) The Wishful-thinking model, which predicts always **low** on the ChargeGroup feature, accuracy on the test set is: **48.88%**

b) Mode of the training ‘chargeGroup’: **medium**.

Mode-based test accuracy: **38.81%**.

This is lower than the wishful-thinking baseline (48.88%) because the test set has a different class distribution than the training set and contains more “low” cases than “medium”, so copying the training mode does not align with the test distribution, whereas always predicting “low” does (in this case).

c) The accuracy values for different depths are given in Figure 6. Best-performing max depth: **4** (first depth reaching the top accuracy).

Highest test accuracy achieved: **94.03%**.

I would pick ‘max depth = 4’ because deeper trees do not improve accuracy. While also not over simplifying.

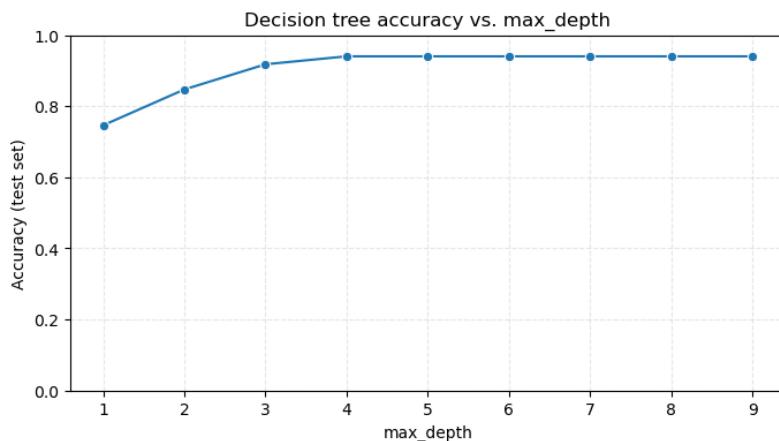


Figure 6: Graph on decision tree accuracy for depths up to 9 on the test dataset

d) The decision tree visualized in Figure 7 classifies chargeGroup as **medium** only when age  $> 44.5$  and when the patient doesn’t smoke ( $\text{smoker\_yes} \leq 0.5$ ).

The specified 42 year old non-smoking male is predicted as **low** because he falls into the age  $\leq 44.5$  branch, and after that into the leaf low because he doesn’t smoke.

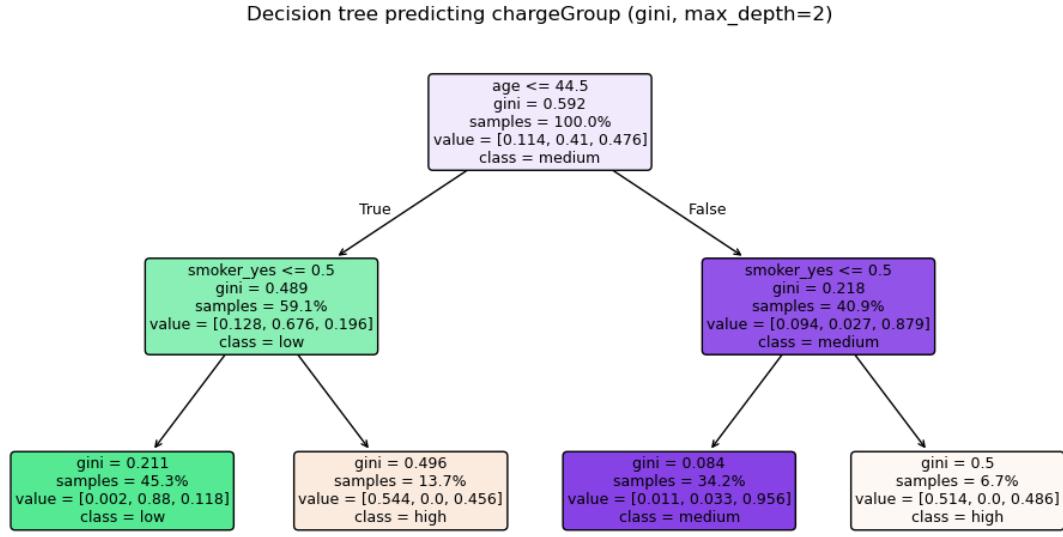


Figure 7: Decision Tree for predicting the chargeGroup based on age and smoking status. The left branch is taken when the decision rule evaluates to true.

### Q3 Clustering

- a) We perform k-means clustering on the dataset. The clusters have the following centroids, sizes and numbers of smokers and non-smokers:

cluster	centroid				size	smoker
	age	bmi	children	charges		yes/no
0	27.38	28.83	0.46	6220.19	498	55/443
1	50.86	32.72	0.55	21060.61	480	157/323
2	40.03	30.47	2.71	12636.33	360	62/298

The clusters are visualized in Figure 8. Cluster 0 is categorized by younger patients with few children, the smallest smoker-percentage and lower charges while cluster 1 consists of older patients that have a slightly higher weight and a higher percentage of smokers and also have few children. The charges for this cluster are spread across the whole spectrum but also include the highest charges that appear in no other cluster. Cluster 2 consists of middle-aged patients with comparatively many children and about average BMI and charges.

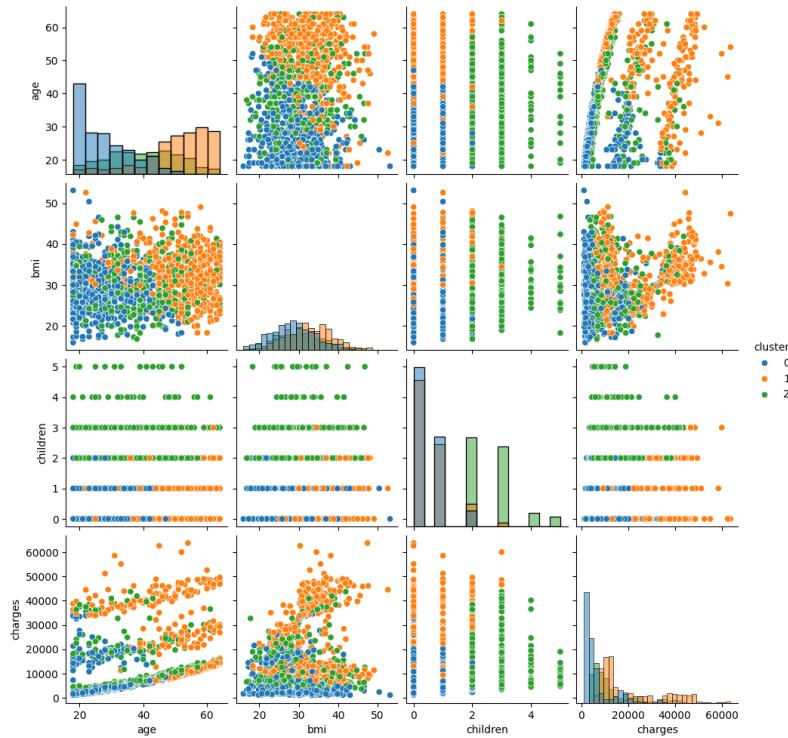


Figure 8: Scatter plot matrix of the numerical features with clusters as hue

## Q4 Regression

a) In Figure 9 the trend line slopes upward, but the scatter remains very wide, so people at the same age can have charges that differ by tens of thousands of dollars. Because smoking status, BMI, and dependents strongly influence charges, a simple linear model on age alone would ignore most variance. Therefore, age-only linear regression would be a poor fit, useful perhaps for spotting the general direction but not for accurate predictions.

b) Train MAE: **\$4,186.98**

Test MAE: **\$4,306.85**.

Linear Regression captures smoking status as the dominant effect (see coefficients in Figure 10), but several other features contribute smaller adjustments, so residuals of several thousand dollars remain. The small gap between train and test MAE ( $\sim \$120$ ) indicates that the model is not severely overfitting. However the MAE is relatively high compared to the charges ranges, so the model should be refined.

## Report Assignment Part 1

---

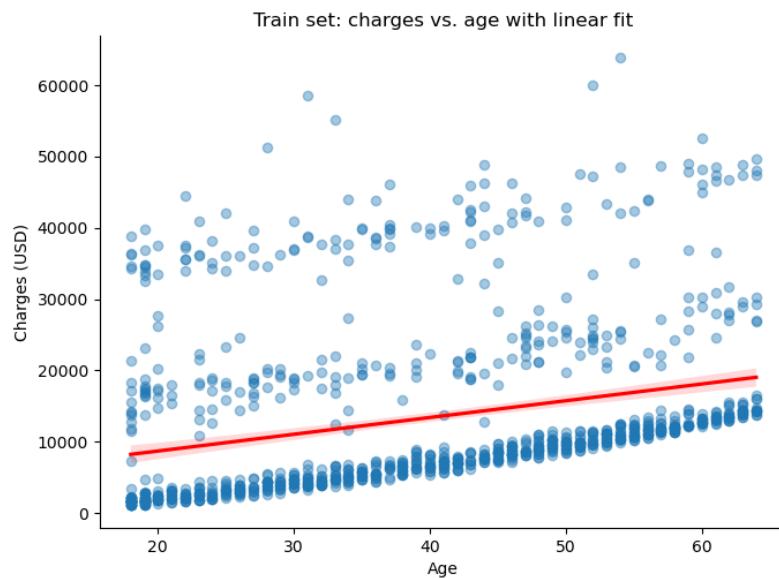


Figure 9: Scatter plot of training data showing charges on the y-axis and age on the x-axis with a fitted linear regression line in red.

	feature	coefficient
0	smoker_no	-11793.39
1	smoker_yes	11793.39
2	region_northeast	622.32
3	region_southeast	-536.06
4	children	376.80
5	bmi	354.41
6	region_southwest	-261.71
7	age	257.53
8	region_northwest	175.45
9	sex_male	-38.44
10	sex_female	38.44

Figure 10: Coefficient list of all features of the linear regression model

## Q5 Support Vector Machines

- a) The confusion matrix of SVM binary classifiers using only the three features age, bmi, and smoker and all descriptive features (except id).can be seen in 11

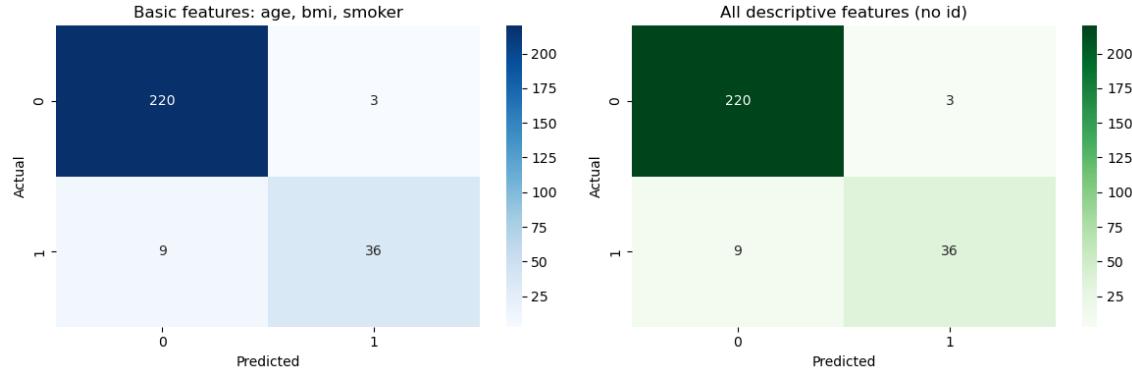


Figure 11: Confusion Matrix

- b) The Basic feature SVM has an accuracy of 0.955 and a precision of 0.923. Same for the All-feature SVM with accuracy: 0.955 and precision: 0.923. So adding other descriptive features doesn't increase the accuracy or precision, which doesn't match to our expectations, since we would have thought that adding information would result in better performance (This would be the assumption without reading ahead to the other tasks).
- c) Basic scaled ( $C=10$ ): accuracy 0.955, precision 0.923, confusion All scaled ( $C=10$ ): accuracy 0.959, precision 0.947, confusion Scaling mainly helped the all-feature model (slightly higher accuracy/precision) by balancing feature magnitudes.

$C$	accuracy
0	10.0
1	5.0
2	1.0
3	0.5
4	0.1

Table 1: Accuracy values for different hyperparameter  $C$ .

- d) How C influences the accuracy: Lower C values enforce a wider margin and tolerate more misclassifications. Conversely, large C emphasises fitting every expensive

example, boosting accuracy but risking overfitting, hence tuning C balances margin width and classification error. In the table 1 we can see that changing the C doesn't greatly impact on the accuracy in this case.

e) Scaling equalizes feature ranges so the hyperplane no longer over weights large-magnitude attributes (e.g., 'age' vs. one-hot binaries). After normalization the decision boundary uses shape rather than scale, which improved both recall and precision in our confusion matrices. Because SVM margins depend on dot products, standardized features lead to more stable optimization and better generalization accuracy.

## Q6 Neural Networks Naive Bayes

a) Baseline MLP accuracy: 0.388 What is striking about the confusion Matrix in Figure 12 is that the model always predicts medium and never high or low. probability variance: 0.02

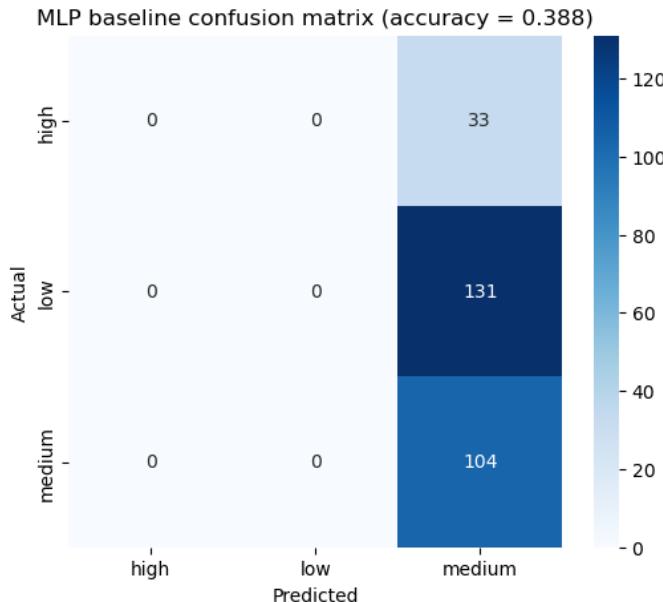


Figure 12: Confusion Matrix of MLP baseline

b) We improved the accuracy by applying the standard scaler beforehand and also removing nodes and having only 3 layers. The confusion matrix of the improved MLP with accuracy of 0.87 can be seen in Figure 13.

## Report Assignment Part 1

---

	$P(\text{high})$	$P(\text{low})$	$P(\text{medium})$
0	0.12	0.41	0.47
1	0.12	0.41	0.47
2	0.12	0.41	0.47
3	0.12	0.41	0.47
4	0.12	0.41	0.47

Table 2: Predicted probability distribution across classes high, low, and medium.

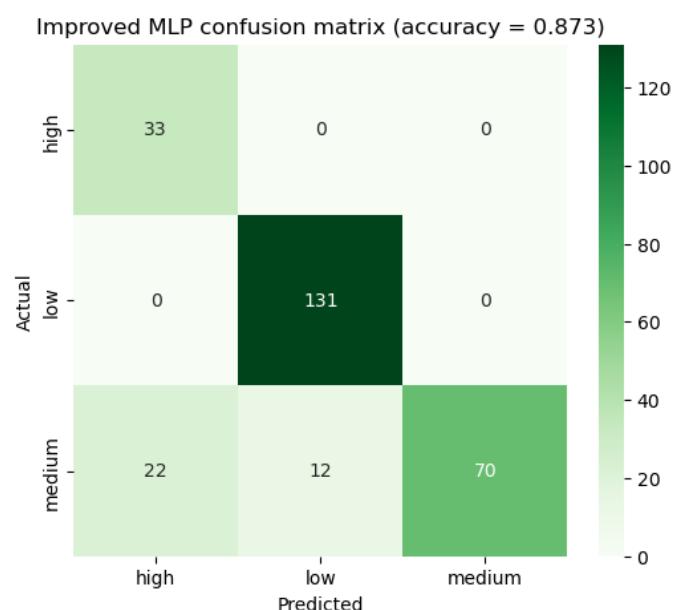


Figure 13: Confusion Matrix of improved MLP

- c) The GaussianNB raw accuracy is 0.62 and the scaled accuracy is 0.62 aswell. the confusion matrices can be seen in Figure 14.

## Report Assignment Part 1

---

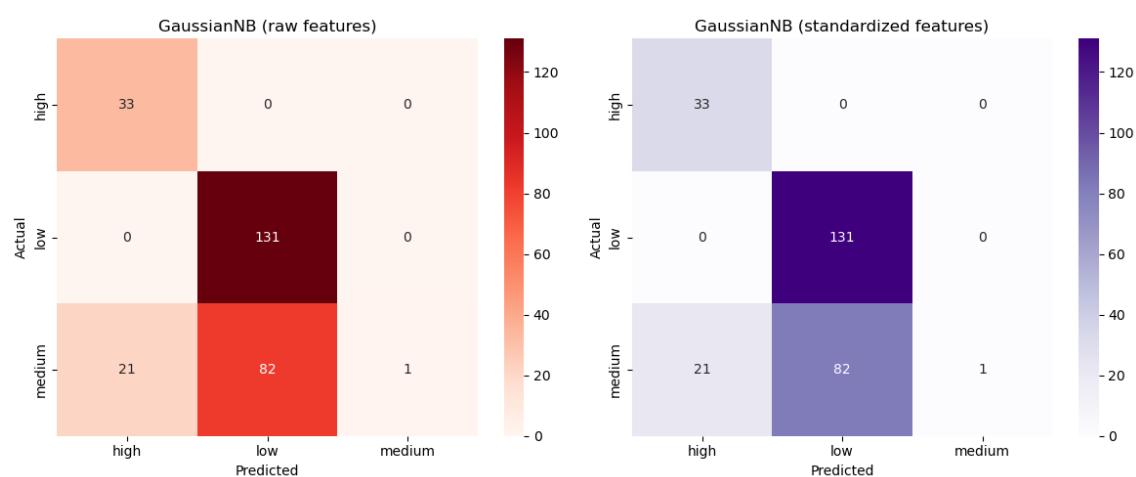


Figure 14: Confusion Matrix of Gaussian Naive Bayes model