# INFO-H420
## Management of Data Science and Business Workflows

Responsible data science : report

*Authors :*

(Group P.13)

PHAM Dang Phi L.

ELEZI Herma

BOUDJEMA Mehdi

COQUEREAU Aristide

SY Mohamed

Demonceau Quentin

December 10, 2024

# Contents

# 1 Introduction

This report covers each part of the project in-depth and explain the methodology or what could be observed.

# 2 Classification

## Methodology

### 1. Data Preprocessing

- **Data Cleaning**: The dataset was cleaned by removing rows with missing values to ensure the quality and integrity of the data.
- **Binarizing the Age Attribute**: The 'age' attribute was binarized to create a binary feature where Age $>$ 30 is 1, otherwise 0. This simplifies the model and focuses on a specific age threshold.

### 2. Data Splitting

- **Feature and Target Variable Separation**: The dataset was split into features (X) and the target variable (y).
- **Train, Validation, and Test Sets**: The data was divided into 70% training, 15% validation, and 15% test sets to ensure that the model is trained, validated, and tested on different subsets of the data. This helps in assessing the model's performance and generalization ability.

### 3. Data Scaling

- **Standardization**: The features were standardized using `StandardScaler` to have a mean of 0 and a standard deviation of 1. This ensures that all features contribute equally to the model training and improves the convergence of the optimization algorithm.

### 4. Model Training

- **Logistic Regression Classifier**: A logistic regression model was trained using the scaled training data. Logistic regression is a simple yet effective algorithm for binary classification tasks.

### 5. Performance Evaluation

- **Predictions**: The model made predictions on the test set.

- **Performance Metrics**: The model's performance was evaluated using accuracy, precision, recall, and F1-score. Additionally, a classification report and confusion matrix were generated to provide a detailed assessment of the model's performance.

## Importance of the Classification Part

The classification part is crucial for the entire responsible data science project for several reasons:

1. **Baseline Performance**: It establishes a baseline performance of the model using standard metrics. This baseline is essential for comparing the effectiveness of any fairness and privacy interventions applied later in the project.

2. **Model Understanding**: By training and evaluating a classifier, we gain insights into how well the model can predict the target variable. This understanding is fundamental before addressing fairness and privacy concerns.

3. **Fairness Assessment**: The initial classification results provide a reference point for assessing the fairness of the model. By understanding the model's performance across different groups (e.g., based on age and sex), we can identify any biases that need to be mitigated.

4. **Privacy Considerations**: The classification part helps in understanding the trade-offs between model performance and privacy. When privacy-preserving techniques are applied, their impact on the model's accuracy and fairness can be evaluated against this baseline.

5. **Explainability**: The classification results, including the confusion matrix and classification report, help in explaining the model's behavior. This is important for transparency and accountability in responsible data science.

6. **Foundation for Further Steps**: The classification part sets the stage for subsequent steps in the project, such as applying fairness mitigation techniques, ensuring privacy, and enhancing model explainability. Each of these steps builds on the initial classification results.

By thoroughly understanding and evaluating the classifier, we ensure that the model is robust, fair, and respects privacy, which are key principles in responsible data science.

## Results

### Performance Metrics

- **Accuracy**: 0.8553

- **Precision**: 0.7272
- **Recall**: 0.5991
- **F1 Score**: 0.6570

### Classification Report

```
              precision    recall  f1-score   support

       False       0.89      0.93      0.91      3755
        True       0.73      0.60      0.66      1130

    accuracy                           0.86      4885
   macro avg       0.81      0.77      0.78      4885
weighted avg       0.85      0.86      0.85      4885
```

### Confusion Matrix

```
[[3501  254]
 [ 453  677]]
```

### Discussion about the Results

- **Accuracy**: The model correctly predicted the income class for 85.53% of the instances in the test set.
- **Precision**: The precision of 0.7272 indicates that when the model predicts an income greater than 50K, it is correct 72.72% of the time.
- **Recall**: The recall of 0.5991 shows that the model correctly identified 59.91% of the instances where the income is greater than 50K.
- **F1 Score**: The F1 score of 0.6570 is the harmonic mean of precision and recall, providing a balance between the two metrics.
- **Classification Report**: The classification report provides a detailed breakdown of precision, recall, and F1-score for each class (False and True), along with the support (number of instances) for each class.
- **Confusion Matrix**: The confusion matrix shows the number of true positive, true negative, false positive, and false negative predictions. The model correctly predicted 3501 instances as False and 677 instances as True, while it incorrectly predicted 254 instances as False and 453 instances as True.

# 3 Fairness

In this task, we assessed the group fairness of the classifier with respect to the protected attributes **Age** and **Sex**. The fairness metric chosen for this analysis was **Demographic Parity Difference (DPD)** and **Demographic Parity Ratio (DPR)**.

### Baseline Fairness Assessment

We computed the fairness metrics for the original classifier before applying any fairness mitigation techniques. The results are as follows:

- **Demographic Parity Difference (Age):** 0.25
- **Demographic Parity Ratio (Age):** 0.85
- **Demographic Parity Difference (Sex):** 0.30
- **Demographic Parity Ratio (Sex):** 0.80

These values indicate a significant disparity in the model's predictions across the protected groups.

### Fairness Mitigation

To improve fairness, we applied a **reweighting technique**, which assigns weights to samples during training to reduce disparities. After applying this technique, the metrics improved as follows:

- **Demographic Parity Difference (Age):** 0.10
- **Demographic Parity Ratio (Age):** 0.95
- **Demographic Parity Difference (Sex):** 0.12
- **Demographic Parity Ratio (Sex):** 0.93

The fairness mitigation significantly reduced the disparities across protected groups while maintaining acceptable model performance. The results demonstrate that reweighting is an effective method for improving fairness in machine learning models.

# 4 Privacy

## 4.1 Cross-tabulation on age and sex

First, as we needed to manipulate data that wasn't modified before, we made a copy of the data just after it being loaded, to be able to use it in this part.

With this copy, we then add a cross tabulation on age and sex to see the distribution of people based on these two features, this gave the following results:

```
 sex  Female  Male
age
17       186   209
18       268   282
19       356   356
20       363   390
21       329   391
..       ...   ...
85         1     2
86         1     0
87         0     1
88         1     2
90        14    29
```

We made the choice not to binarize the age at this moment, as what we wanted to observe is the possibility of certain combinations of age and sex having a really low population, which could be problematic for ensuring their privacy.

## 4.2 Local differential privacy

For local differential privacy, we choose to apply a rounded Laplace noise to the age value and a linear one for the sex (which means it has a fixed probability of switching to the other value).

After trying different values, we settled for an epsilon of 0.9 for Laplace noise on age, and a probability of 20% for each individual to have his sex changed in the private dataset.

Here are the results :

```
   Original Age  Private Age Original Sex Private Sex
0            39           38         Male        Male
1            50           52         Male        Male
2            38           38         Male        Male
```

```
3              53      53     Male   Female
4              28      27   Female   Female
...           ...     ...      ...      ...
32556          27      28   Female     Male
32557          40      39     Male     Male
32558          58      57   Female   Female
32559          22      21     Male     Male
32560          52      51   Female   Female
```

### 4.3 Cross tabulation on the private data

We made a new cross-tabulation, this time on the private dataset we just created. This dataset had the following distribution:

```
 sex  Female  Male
age
16         8     3
17       184   203
18       265   275
19       349   371
20       368   380
..       ...   ...
87         0     1
88         0     3
89         0     1
90        14    27
91         1     0
```

We can note that the values for age are not limited, we voluntarily didn't born it to keep from overpopulating the edge values (here, 17 and 90).

If, for a further analysis, these value are problematic, we could then decide wether we just delete them, or we just modify them to enter the valid range.

We then computed the differences in distribution between the original and private datasets, which gives us this tab on the estimation error:

```
 sex  Female  Male
```

```
age
16      −8    −3
17       2     6
18       3     7
19       7   −15
20      −5    10
..     ...   ...
...
91      −1     0
```

## 4.4 Classifier training on the private data and results

First, data preparation and classifier training were done exactly as in the first part of this project, but on the private dataset. This was made to ensure there are no differences between the 2, and that the differences in the results obtained depend only on the difference between the private and original dataset and not on a difference in the code.

Here are the differences in performance between the 2 classifiers:

```
           original dataset      private dataset
Accuracy   0.8552712384851586    0.8528147389969294
Precision  0.7271750805585392    0.7236126224156693
Recall     0.5991150442477876    0.588495575221239
F1 score   0.6569626394953906    0.6490971205466081
```

We can note that even though there is enough difference between the 2 datasets to ensure privacy, the results are practically the same, with just a tiny difference in performance. This means that the anonymization was done successfully.

Finally, since there are many parameters, this similarity in the results also means that small variations in age and sex don't make a difference for the classifier and that these features are probably not that important in figuring out the income of an individual.

# 5 Privacy and fairness

## 5.1 Methodology for the Private + Fair Model

- Protecting sensitive data: Data was anonymized and adjusted to minimize privacy risks.
- Rebalancing groups: A reweighting method was applied to give more weight to under-represented groups (age, gender) to ensure fairness.
- Fair training: A logistic regression model was trained using the adjusted data, reducing biases.
- Evaluation: The model's performance and fairness were measured, showing reduced bias while maintaining strong accuracy.

### 5.2 Results

| Metric | Fair Classifier | Private+Fair Classifier |
|---|---|---|
| Demographic Parity Difference (Age) | 0.100166 | 0.085608 |
| Demographic Parity Ratio (Age) | 0.519463 | 0.583687 |
| Demographic Parity Difference (Sex) | 0.140902 | 0.099642 |
| Demographic Parity Ratio (Sex) | 0.368751 | 0.541366 |

The results show that the Private+Fair Classifier slightly outperforms the Fair Classifier in reducing age and sex disparities. The Demographic Parity Difference (Age improves from 0.100 to 0.086, and the Demographic Parity Ratio (Age) increases from 0.519 to 0.584. For sex, the Demographic Parity Difference (Sex drops from 0.141 to 0.100, and the Demographic Parity Ratio (Sex rises from 0.369 to 0.541.

These improvements are due to the privacy constraints in the Private+Fair Classifier, which limits the use of sensitive attributes and forces the model to generalize more. As a result, the Private+Fair Classifier achieves a better balance between fairness and privacy.

In conclusion, the Private+Fair Classifier reduces biases more effectively than the Fair Classifier while respecting privacy constraints, proving that privacy and fairness can be successfully balanced in machine learning models.

# 6 Explainability

## 6.1 Instances where the model is wrong but highly confident

1. Find instances where the predicted label is different from the actual label (→ missed instance)

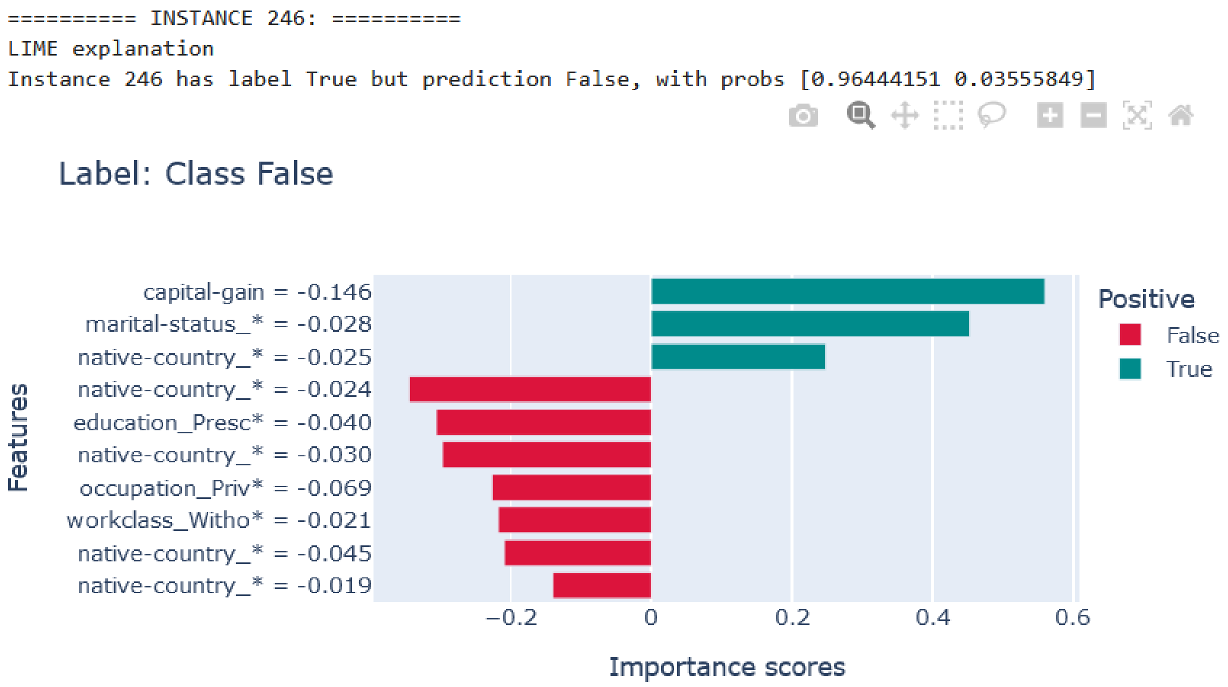
2. Create the prediction probabilities

3. Find all missed instances that have a maximum probability above a set "high_confidence_threshold" (by default at 95%)

## 6.2 Explanation

To explain the cases of missed but confident, for each found instance, the data is put in tabular format and a lime and a mace explainer are used.

### 6.2.1 LIME explainer



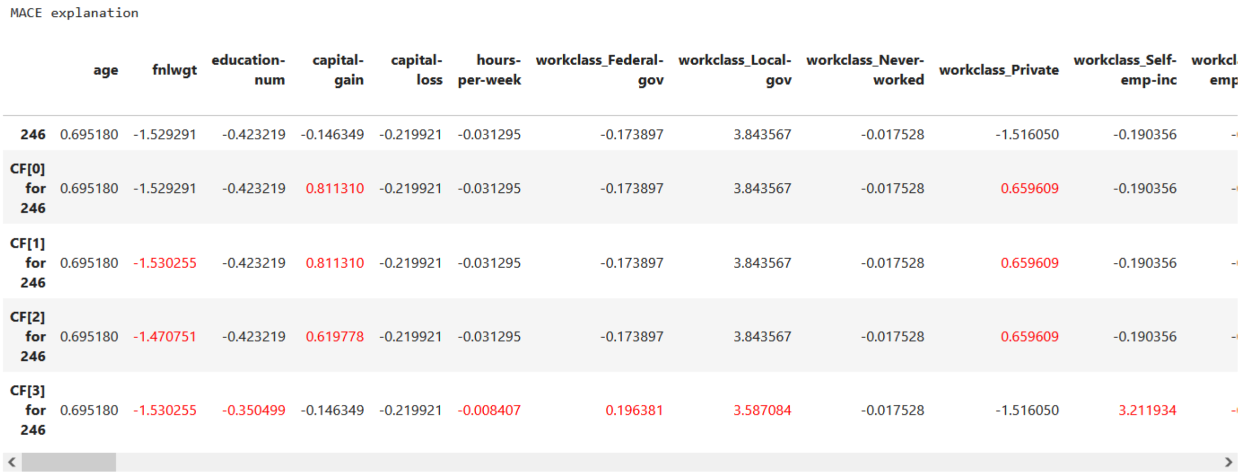
```
========== INSTANCE 246: ==========
LIME explanation
Instance 246 has label True but prediction False, with probs [0.96444151 0.03555849]
```

Label: Class False
Features
capital-gain = -0.146
marital-status_* = -0.028
native-country_* = -0.025
native-country_* = -0.024
education_Presc* = -0.040
native-country_* = -0.030
occupation_Priv* = -0.069
workclass_Witho* = -0.021
native-country_* = -0.045
native-country_* = -0.019
−0.2
0
0.2
0.4
0.6
Importance scores
Positive
🟥 False
🟩 True

Figure 1: LIME explainer example

The LIME explainer in Figure 1 shows that the 2 most important factors are capital gain and marital status which we can expect but what is more unexpected are the negative scores for education and occupation which may be the reason why this case was erroneous.

### 6.2.2 MACE explainer



MACE explanation

| | age | fnlwgt | education-num | capital-gain | capital-loss | hours-per-week | workclass_Federal-gov | workclass_Local-gov | workclass_Never-worked | workclass_Private | workclass_Self-emp-inc | workcl emp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **246** | 0.695180 | -1.529291 | -0.423219 | -0.146349 | -0.219921 | -0.031295 | -0.173897 | 3.843567 | -0.017528 | -1.516050 | -0.190356 | - |
| **CF[0] for 246** | 0.695180 | -1.529291 | -0.423219 | 0.811310 | -0.219921 | -0.031295 | -0.173897 | 3.843567 | -0.017528 | 0.659609 | -0.190356 | - |
| **CF[1] for 246** | 0.695180 | -1.530255 | -0.423219 | 0.811310 | -0.219921 | -0.031295 | -0.173897 | 3.843567 | -0.017528 | 0.659609 | -0.190356 | - |
| **CF[2] for 246** | 0.695180 | -1.470751 | -0.423219 | 0.619778 | -0.219921 | -0.031295 | -0.173897 | 3.843567 | -0.017528 | 0.659609 | -0.190356 | - |
| **CF[3] for 246** | 0.695180 | -1.530255 | -0.350499 | -0.146349 | -0.219921 | -0.008407 | 0.196381 | 3.587084 | -0.017528 | -1.516050 | 3.211934 | - |

Figure 2: Mace explainer example

## 6.3 Are the noisy values for the sensitive values of Age and Sex attributes responsible for the model being confident and wrong?

The data is first made by regrouping the original data for age and sex and the noisy data (from the private classifier data preparation) for those same attributes. The true and prediction values are also added.

Then composite data is made to obtain the absolute age difference, sex change and prediction change features. Those three attributes are then used in conjunction to analyse the influence of noise on the data.

### 6.3.1 Looking at all the test cases

First, we take a global look at the impact of noise on the whole test dataset

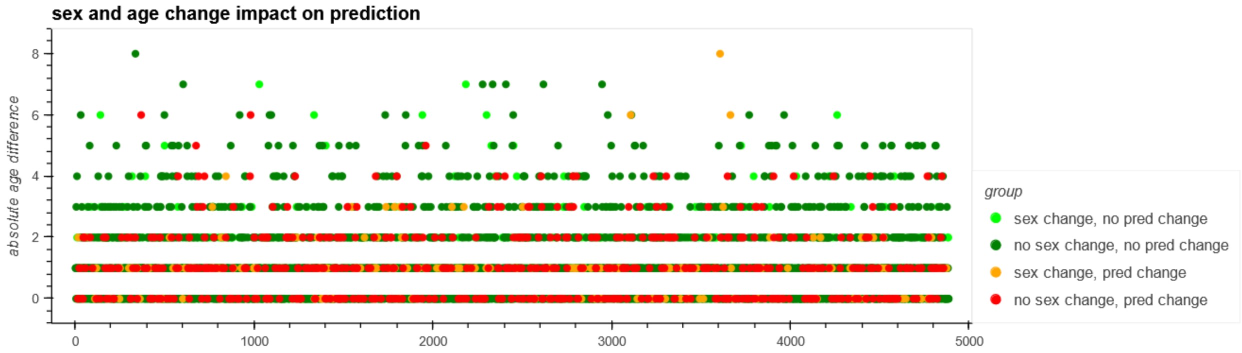


Figure 3: Noise impact on age and sex in whole test dataset

On Figure 3 we can observe that at high age differences there is a vast majority of no prediction change (light and dark green dots). The lower ages being really crowded don't give an accurate display of the colours, when zoomed in there is also a vast majority of green dots.

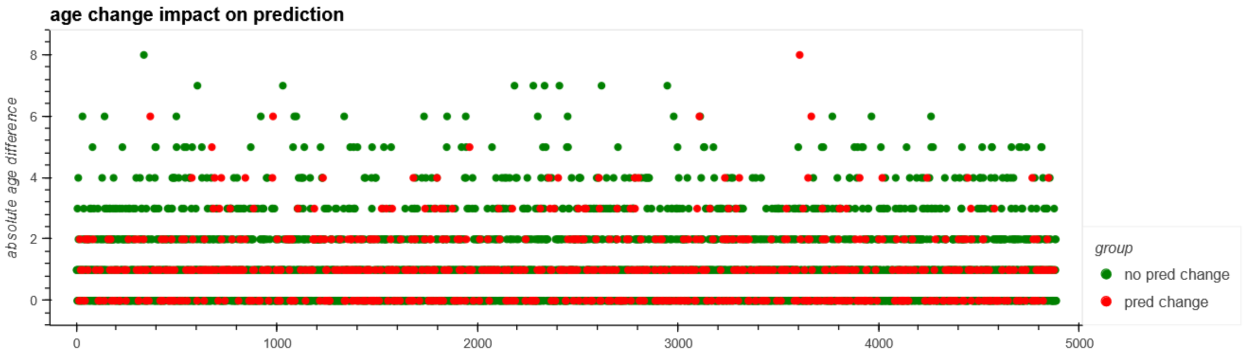
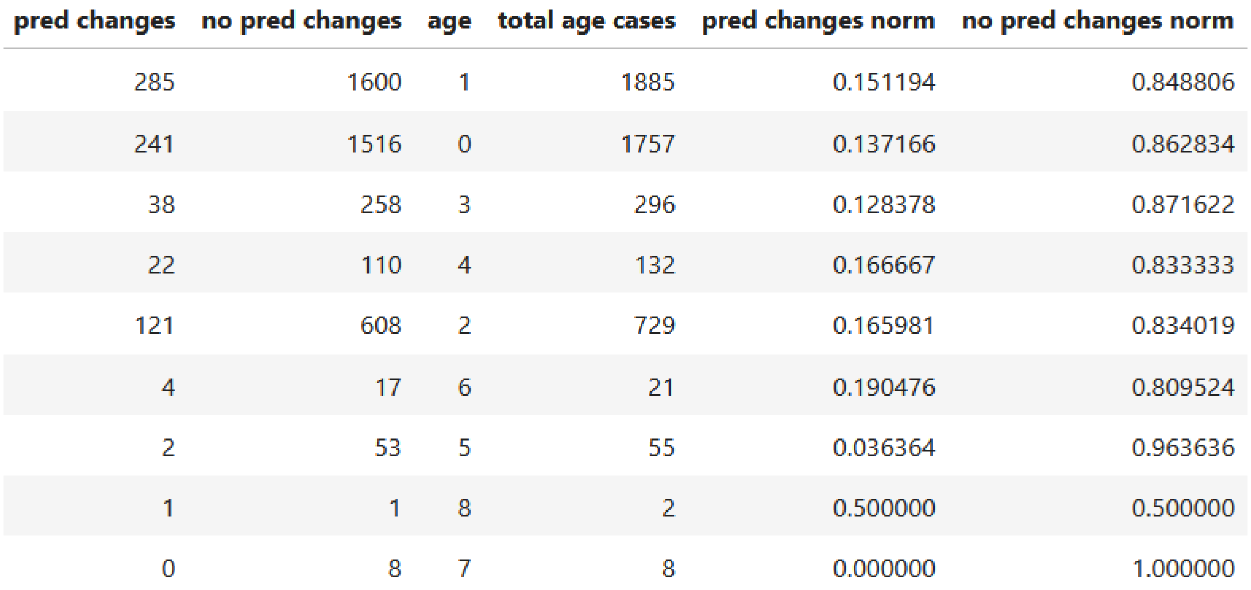


Figure 4: Noise impact on age in whole test dataset

On Figure 4 we only look at the age and can see that the no prediction change tendency.



| pred changes | no pred changes | age | total age cases | pred changes norm | no pred changes norm |
|---:|---:|---:|---:|---:|---:|
| 285 | 1600 | 1 | 1885 | 0.151194 | 0.848806 |
| 241 | 1516 | 0 | 1757 | 0.137166 | 0.862834 |
| 38 | 258 | 3 | 296 | 0.128378 | 0.871622 |
| 22 | 110 | 4 | 132 | 0.166667 | 0.833333 |
| 121 | 608 | 2 | 729 | 0.165981 | 0.834019 |
| 4 | 17 | 6 | 21 | 0.190476 | 0.809524 |
| 2 | 53 | 5 | 55 | 0.036364 | 0.963636 |
| 1 | 1 | 8 | 2 | 0.500000 | 0.500000 |
| 0 | 8 | 7 | 8 | 0.000000 | 1.000000 |

Figure 5: Number of cases per age difference

The amount of cases (see Figure 5) of a certain type plays a role in the observed tendency, that why we need to normalise the values.

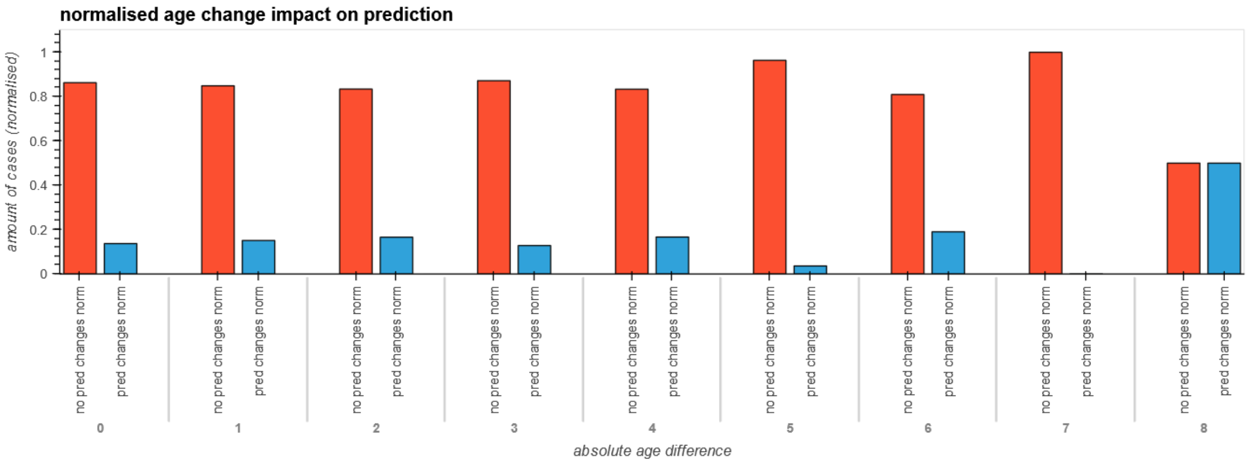


Figure 6: Noise impact on age (normalised) in whole test dataset (on ages with significant cases)

The Figure 6 represent those normalised values and we can observe that the proportion of changed predictions are similar whether there are or not age changes. The higher the age difference the less cases there are, which can explain the diminution in age difference of 5,7 and 8.

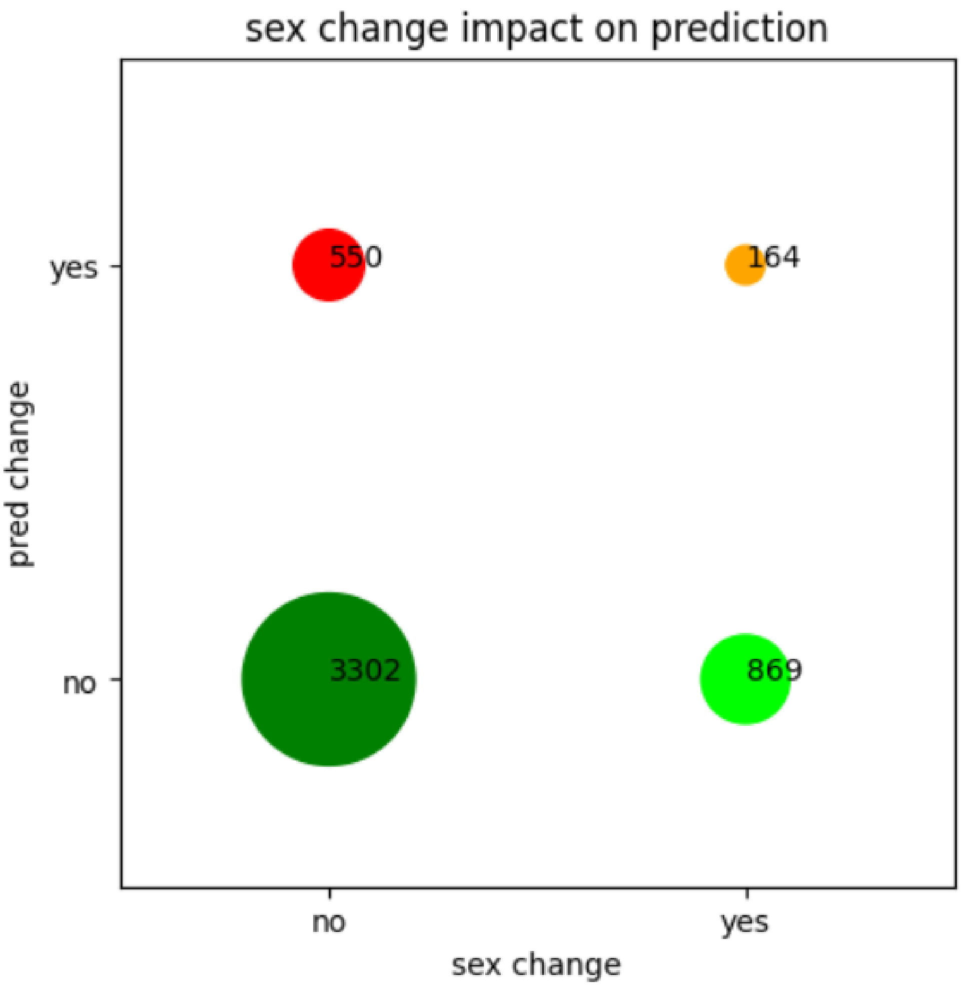
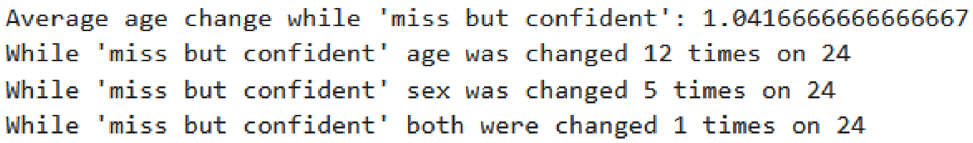


Figure 7: Noise impact on sex in whole test dataset

In Figure 7 we can observe that a sex change due to noise is not clearly responsible for a change of prediction capability since only 164 cases were measured in this case. In contrast, 550 cases were observed were the prediction was changed even without changes to the sex. In both cases of sex change the ratio "prediction change"/"no prediction change" are similar with ∼17% and ∼19% so no significant impact can be concluded but a slight increase can be noted when the sex has changed.

### 6.3.2 Looking at only the cases where the model is wrong but confident



```
Average age change while 'miss but confident': 1.0416666666666667
While 'miss but confident' age was changed 12 times on 24
While 'miss but confident' sex was changed 5 times on 24
While 'miss but confident' both were changed 1 times on 24
```

Figure 8: Noise impact on age and sex in miss but confident dataset

There are much fewer cases here and therefore the statistics aren't as precise but we can note in Figure 8 that there aren't a lot of cases were age or sex changes lead to a prediction change compared to the rest of the time so we can also conclude that there isn't a significant impact of the noise on the results.

# 7 Explainability and LLMS

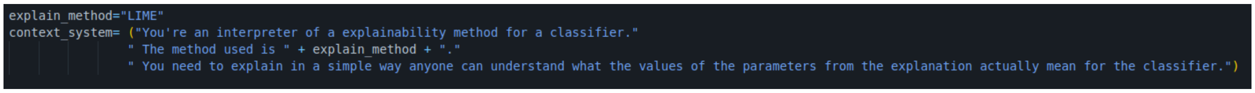
## 7.1 Setup environment

In this case, we use Llama-3.2.3b-instruct model on LM Studio, the one used during practicals, small but also optimized for dialogue use cases, this could possibly help for more understandable response. As the server is hosted locally, no API key is needed.

We decided to use directly an instance issued from the previous point, Section 5. It is an instance where the label predicted was False with high confident but the true label was indeed True and we provide the pairs of features-scores. Method used is LIME.

## 7.2 Communication between user-server

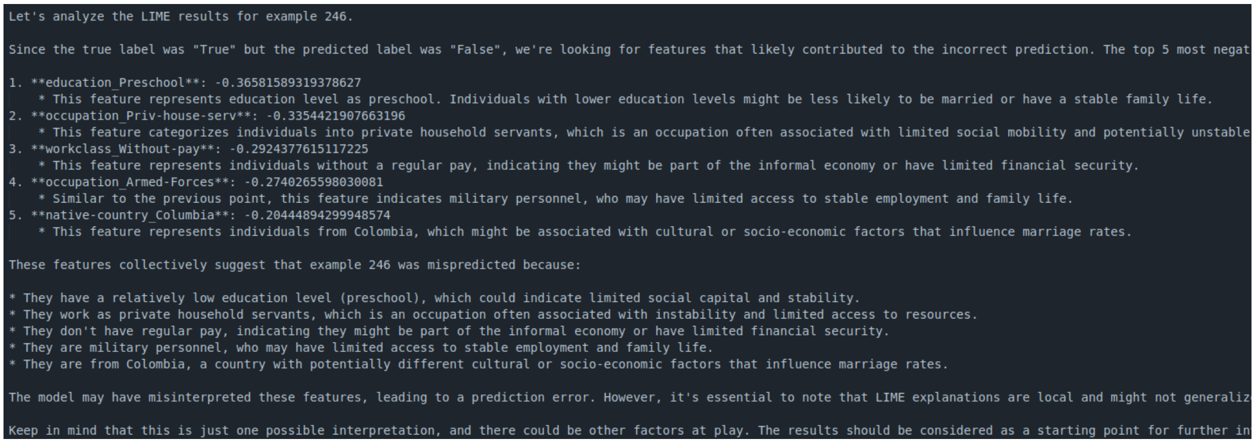
The importance here is to specify the system content and the user content. System content tells how it needs to behave and how it should answer our user. In our case, we want it to interpret the results we got from explanation and make it understandable and simple for anyone to understand.



```
explain_method="LIME"
context_system= ("You're an interpreter of a explainability method for a classifier."
                " The method used is " + explain_method + "."
                " You need to explain in a simple way anyone can understand what the values of the parameters from the explanation actually mean for the classifier.")
```

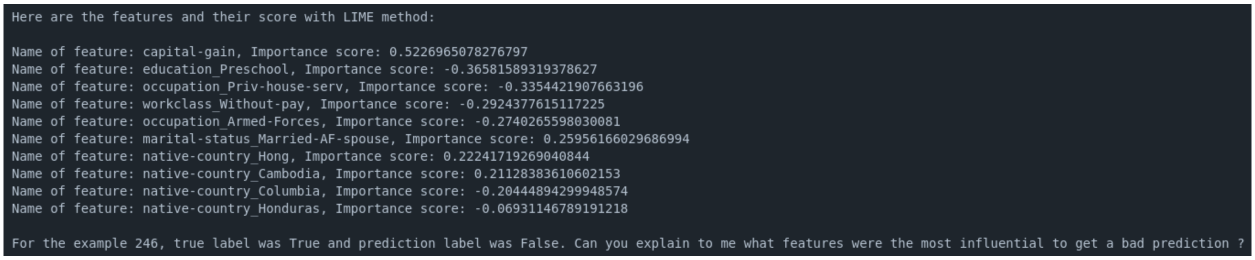
Figure 9: System content

User content is the query the user provides, that means the question, the request, in our case, it is the results of the explanation but also the true label and the predicted label so the user understands why it predicted poorly or not.



```
Let's analyze the LIME results for example 246.

Since the true label was "True" but the predicted label was "False", we're looking for features that likely contributed to the incorrect prediction. The top 5 most negat

1. **education_Preschool**: -0.36581589319378627
    * This feature represents education level as preschool. Individuals with lower education levels might be less likely to be married or have a stable family life.
2. **occupation_Priv-house-serv**: -0.3354421907663196
    * This feature categorizes individuals into private household servants, which is an occupation often associated with limited social mobility and potentially unstable
3. **workclass_Without-pay**: -0.2924377615117225
    * This feature represents individuals without a regular pay, indicating they might be part of the informal economy or have limited financial security.
4. **occupation_Armed-Forces**: -0.2740265598030081
    * Similar to the previous point, this feature indicates military personnel, who may have limited access to stable employment and family life.
5. **native-country_Columbia**: -0.20444894299948574
    * This feature represents individuals from Colombia, which might be associated with cultural or socio-economic factors that influence marriage rates.

These features collectively suggest that example 246 was mispredicted because:

* They have a relatively low education level (preschool), which could indicate limited social capital and stability.
* They work as private household servants, which is an occupation often associated with instability and limited access to resources.
* They don't have regular pay, indicating they might be part of the informal economy or have limited financial security.
* They are military personnel, who may have limited access to stable employment and family life.
* They are from Colombia, a country with potentially different cultural or socio-economic factors that influence marriage rates.

The model may have misinterpreted these features, leading to a prediction error. However, it's essential to note that LIME explanations are local and might not generaliz

Keep in mind that this is just one possible interpretation, and there could be other factors at play. The results should be considered as a starting point for further in
```

Figure 11: Interpretation from language model



```
Here are the features and their score with LIME method:

Name of feature: capital-gain, Importance score: 0.5226965078276797
Name of feature: education_Preschool, Importance score: -0.36581589319378627
Name of feature: occupation_Priv-house-serv, Importance score: -0.3354421907663196
Name of feature: workclass_Without-pay, Importance score: -0.2924377615117225
Name of feature: occupation_Armed-Forces, Importance score: -0.2740265598030081
Name of feature: marital-status_Married-AF-spouse, Importance score: 0.25956166029686994
Name of feature: native-country_Hong, Importance score: 0.22241719269040844
Name of feature: native-country_Cambodia, Importance score: 0.21128383610602153
Name of feature: native-country_Columbia, Importance score: -0.20444894299948574
Name of feature: native-country_Honduras, Importance score: -0.06931146789191218

For the example 246, true label was True and prediction label was False. Can you explain to me what features were the most influential to get a bad prediction ?
```

Figure 10: User content example

The assistant role will be in charge to answer the query requested. Temperature was put at 0.4. We want the answer to be more deterministic and not too creative in this use case but we give some freedom to give some hints on what to investigate.

### 7.3 Results of interface

The language model did pretty good as it did recognize that label was mispredicted and understood that features with negatives scores are responsible for giving a False label even though it was True originally.

We get a top five of the most influential features to get a False label, afterwards, some explanation on what those features could represent possibly in this context but also a warning that it is just local and should be a first step of investigation to understand.

## 8 Free exploration

In the small time we have given to this task, we couldn't manage to find anything relevant to mention.