This document explains what is expected of you during the realization of your project. You will have three main renderings:

* **report 1** : introduce, describe, visualize and manipulate your dataset;
* **report 2** : modeling and solving the problem using Machine Learning techniques;
* **final report** : made up of the first two reports, it embodies your vision and your work in relation to your project by integrating a conclusion and an opening.

**This methodology is therefore intended to advise you in the production of renderings, but does not represent specific instructions that you must follow. You are in control of your work, your analyses, as well as the format that your various reports will take.**

**Report 1: exploration, data visualization and data pre-processing report**

**Introduction to the project**

**Context**

* Context of the project's integration into your business.
* From a technical point of view.
* From an economic point of view.
* From a scientific point of view.

The Heartbeat Analysis project is an international collaboration between Ping Yuan (France), Dr. Sepideh Tabrik (Germany), and Isabell Gurstein (Germany). As we come from different business contexts, we can only answer the question of business integration of the project individually.

Ping Yuan: I work in Datategy which is one startup for data scientist daily tools. With about 8 years in development, I have rich experience on implement features in code. I am interesting to the data scientist topics. My goal is to gain my knowledge of mathematics to help me improve the understanding and apply the data scientist concept in the daily life. Hope I can have the choice to switch my career path to data scientist after this classes.

Dr Sepideh Tabrik: …

Isabell Gurstein: I work as an SAP BI/Data Analyst at an SAP consulting company in Germany specializing in the retail sector. My goal is to train as a data scientist and subsequently as a machine learning engineer to acquire a similar position within my current company. As part of the data science training, the current project focuses more on the healthcare sector, which has little to do with my current work reality but aligns with my original background.

I chose this project because it provides a unique methodological opportunity to examine sounds as an object of analysis using deep learning models, enabling me to learn the complementary expertise to implement them within my company. Furthermore, with my knowledge, I aim to contribute to broaden the spectrum of customers both in content and methodology in my company.

**Objectives**

The main objectives of the project are:

* Designing, training, and optimizing deep learning models for analyzing heartbeat sounds using deep learning models for classification.
* Providing team members with the opportunity to learn and apply deep learning techniques, equipping them with skills to implement these technologies in their own working fields.
* Facilitating international collaboration to leverage diverse expertise, enhancing the project's outcomes.

Ping Yuan: I have built one frontend to visualize the heartbeat to help medicine make their analysis. But there is no ai or tools to help predict on it. I hope with this project, we can build a docker image which contains the model we built and a server and ui so that the medicine can load the data with the ui and get help on diagnose.

Dr Sepideh Tabrik: …

Isabell Gurstein: I have worked for several years in the field of medical invoice auditing, two years as a research associate in gerontopsychiatric clinical trials, and an additional two years in IT project management at a health insurance company in Germany. Consequently, I have acquired expertise in the heathcare field and statistical analysis. As part of my Master of Science degree, I have also completed several courses in Machine Learning, but I would still classify myself as a late beginner to early intermediate in terms of machine learning proficiency. I currently have only one Machine Learning expert on my team, which strongly focusses on Chatbot technology implementation. Whom I have not contacted regarding the project so far.

* What are the main objectives to be achieved? Describe in a few lines.
* For each member of the group, specify the level of expertise around the problem addressed?
* Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.
* (Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?).

**Understanding and manipulation of data**

**Framework**

* Which set(s) of data(s) did you use to achieve the objectives of your project?
* Are these data freely available? If not, who owns the data?
* Describe the volume of your dataset?

We simultaneously worked on two ECG Heartbeat Categorization Datasets obtained and cleaned beforehand from Kaggle.com for analysis. The first dataset originated from the MIT-BIH Arrhythmia Database (mitbih), while the second was based on the PTB Diagnostic ECG Database (ptbdb). Both the original and cleaned datasets are freely accessible, facilitating easy training and testing of our model on the prepared datasets and validation on the raw ones.

The mitbih dataset on Kaggle.com has 188 columns, where 187 are heartbeat data points represented as heartbeat frequencies in float format, with the last column representing the heartbeat classification. The ptbdb dataset on Kaggle.com has 187 columns, where 186 are heartbeat data points represented similarly, with the last column also representing the heartbeat classification.

The mitbih differs from ptbdb in that it features a multiclass target variable, whereas the PTBDB dataset uses dummy coding for the heartbeat target variable. The target variable in the MIT dataset consists of classes distributed among a normal heartbeat class and four distinct ones ('Supraventricular', 'Premature', 'Fusion', 'Unclassifiable'). In contrast, the ptbdb dataset distinguishes solely between normal and abnormal classes using a dummy variable.

Furthermore, the cleaned data differs from the raw data as it contains fewer variables representing the frequency span for each heartbeat, thereby reducing the amount of information available for each heartbeat class.

Kaggle.com provides both cleaned datasets partitioned into train and test datasets. The mitbih train data comprises 87,554 cases (~75%), and the test data comprises 21,891 cases (~25%).

The ptbdb train data has XXXX (~75%) and the test data has XXXX (~25%) cases. …

## **Relevance**

The relevance of classes in the MIT dataset varies; typically, the 'Unclassifiable' beat does not contribute additional information and complicates classification. Therefore, we chose to exclude this particular class from the modeling process.

Since all explanatory variables are sequential and represent specific parts of the heartbeat sequence for each case, they are equally meaningful in predicting the target variable. Thus, none of the features can be singled out as particularly significant.

* Which variables seem most relevant to you with regard to your objectives?
* What is the target variable?
* What features of your dataset can you highlight?
* Are you limited by some of your data?

## **Pre-processing and feature engineering**

Since the initial data was clean and all features were on the same scale, we performed minimal feature engineering (converted the target variable type from float to numeric and added labels since the data came without them), allowing us to quickly proceed with the preprocessing step.

However, exploratory data analysis revealed that the classes of the target variable were not clearly separated. Specifically, 'Fusion' (depicted in red) represents a mixed class of 'Supraventricular' (yellow) and 'Premature' (orange), while 'Unclassifiable' (gray) could potentially include elements of all other classes. As illustrated in the graph, this complicates the distinction between the signals. Ein Bild, das Text, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Nevertheless, when running histograms over all data points in the mitbih training dataset by class, we realized that the shapes of heart beat classes do indeed differ from each other, as shown in the example of normal and supraventricular heart beats.

Ein Bild, das Screenshot, Farbigkeit, Electric Blue (Farbe), Blau enthält.

Automatisch generierte Beschreibung

Ein Bild, das Screenshot, Text, Electric Blue (Farbe), Blau enthält.

Automatisch generierte Beschreibung

To reduce data complexity, we conducted Principal Component Analysis (PCA) during the exploratory data analysis phase. The Scree Plot indicated that three to four dimensions effectively represent the data.Ein Bild, das Text, Reihe, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Farbigkeit enthält.

Automatisch generierte Beschreibung

Since PCA captures only linear relationships, we additionally applied t-SNE and UMAP to incorporate non-linear dimension reduction methods. Both additional methods emphasized a strong overlap among the groupsEin Bild, das Karte, Text, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung Ein Bild, das Text, Zeichnung, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung

Furthermore, we realized that the data was heavily unbalanced, favoring the normal heartbeat class in both cleaned datasets. For instance, in the mitbih test dataset, the normal class is represented by 72,471 cases (~83%), while all other classes combined are represented by 15,083 cases (~17%).Ein Bild, das Text, Screenshot, Diagramm, Kreis enthält.

Automatisch generierte Beschreibung

Hence, we also applied strategies like … to balance out the classes.

We also applied methods of dimensionality reduction to our datasets …

* Did you have to clean and process the data? If yes, describe your treatment process.
* Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
* Are you considering dimension reduction techniques in the modeling part? If yes, why?

## **Visualizations and Statistics**

* Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
* Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
* Present the statistical analyzes used to confirm the information present on the graphs.
* Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

**Assessment methods:**

**Reconstituted professional situation: from a set of company data, the candidate must implement various pre-processing and data augmentation to make them usable through machine learning techniques.**

**Report 2: modeling report**

# **Stages of the project**

# **Classification of the problem**

* What kind of machine learning problem is your project like? (classification, regression, clustering, etc)
* What task does your project relate to? (fraud detection, facial recognition, sentiment analysis, etc)?
* What is the main performance metric used to compare your models? Why this one?
* Did you use other qualitative or quantitative performance metrics? If yes, detail it.

## **Model choice and optimization**

* What algorithms have you tried?
* Describe which one(s) you selected and why?
* Did you use parameter optimization techniques such as Grid Search and Cross Validation?
* Have you tested advanced models? Bagging, Boosting, Deep Learning… Why?

## **Interpretation of results**

* Have you analyzed the errors in your model?
* Did this contribute to his improvement? If yes, describe.
* Have you used interpretability techniques such as SHAP, LIME, Skater… (Grad-CAM for Deep Learning…)
* What has (or not) generated a significant improvement in your performance?

**Assessment methods:**

**Professional scenario: based on a proposed solution, the candidate will have to produce a summary report including: the explanation of the choices of AI solutions implemented, the interpretation of the results, the evaluation of the reliability of the algorithms and an optimization proposal.**

**Final report :**

# **Conclusion drawn**

# **Difficulties encountered during the project**

* What was the main scientific obstacle encountered during this project?
* For each of the following points, if you encountered difficulties, detail how they slowed you down in setting up your project.
* Forecast: tasks that took longer than expected, etc.
* Datasets: acquisition, volumetry, processing, aggregation, etc.
* Technical/theoretical skills: timing of skill acquisition, skill not offered in training, etc.
* Relevance: of the approach, model, data, etc.
* IT: storage power, computational power, etc.
* Other

# **Report**

* Detail what was your main contribution to achieving the project's goals.
* Have you changed the model since the last iteration? If yes, provide details.
* Present the results obtained and compare them to the benchmark
* For each of the project's goals, detail how they were achieved or not.
* If they have been reached, in which process(es) can your model fit? Detail.

**Continuation of the project**

* What avenues for improvement do you suggest to increase the performance of your model?
* How has your project contributed to an increase in scientific knowledge?

## **Bibliography**

* What bibliographical elements (research articles, blog, books, etc.) did you rely on to carry out your project?

# **Appendices**

* Gantt diagram.
* Description of code files.