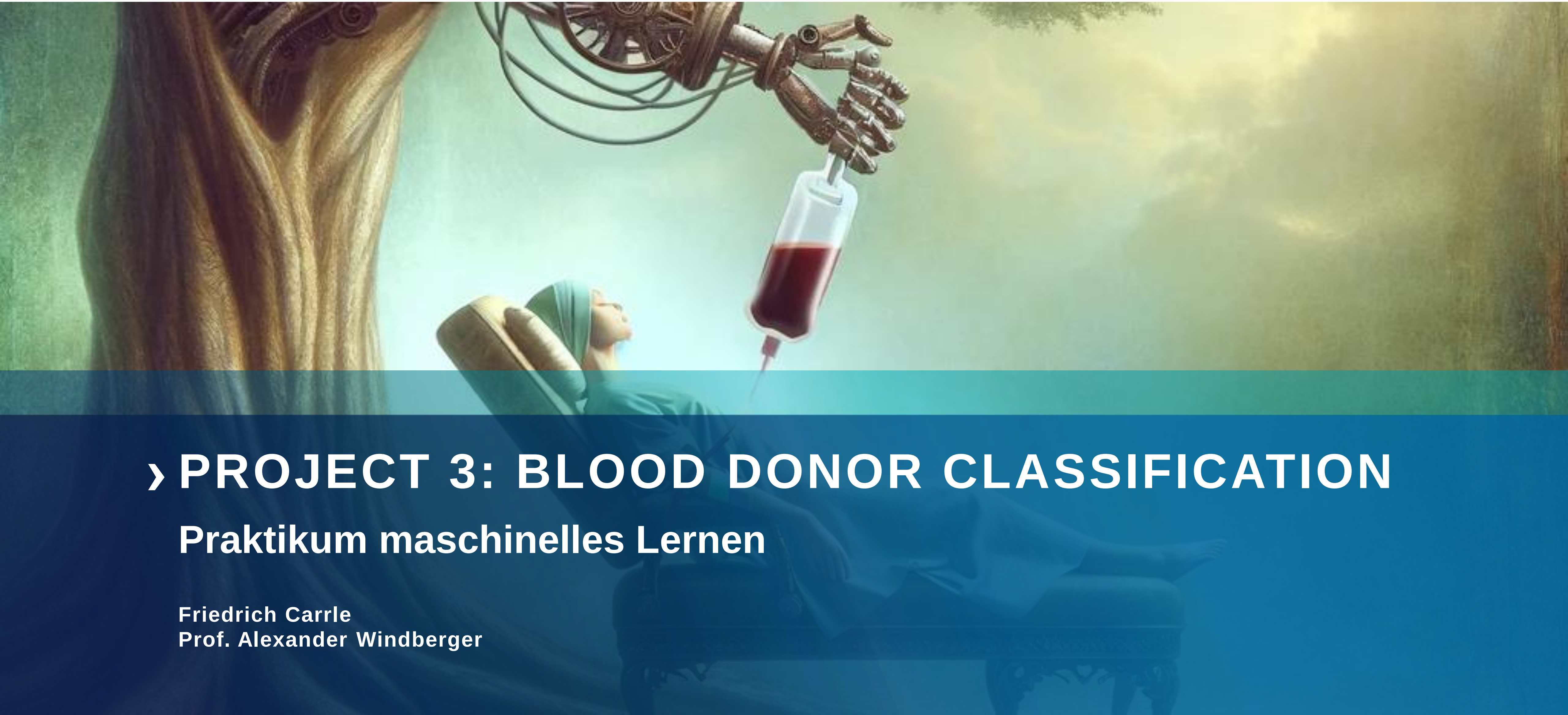
HEILBRONN UNIVERSITY
OF APPLIED SCIENCES



› PROJECT 3: BLOOD DONOR CLASSIFICATION

Praktikum maschinelles Lernen

Friedrich Carrle
Prof. Alexander Windberger

PROJECT 3:
BLOOD DONOR CLASSIFICATION

Develop and evaluate a machine learning pipeline that helps hospital staff to decide whether a person can be a blood donor or if they have medical condition preventing them from donating. The decision is based on hemogram data of several patients (**hemodat.csv**). However, patients with medical conditions are rare and hard to label.

PROJECT 3:
BLOOD DONOR CLASSIFICATION

0. Preperation: First, get to know the dataset and deal with missing values.

- Perform an exploratory data analysis to get to know the data set
- Preprocess the data. If there are missing values, **impute them**.
- **Estimate the accuracy** of your imputation for each feature.

PROJECT 3:
BLOOD DONOR CLASSIFICATION

1. Anomaly Detection: Since medical conditions that lead to the rejection of a donor are rare (luckily) and can be very versatile. It is near impossible to categorize every possible condition. Hence, it would be useful to have an anomaly detection algorithm in place as a safety mechanism to detect suspicious blood samples for further testing.

- Train **an anomaly detection** model based only on valid blood donors without a medical condition.
- **Evaluate the accuracy** of your anomaly detection by testing it also on donors with a medical condition.
- **Perform a PCA** to visualize the true / false positive and true / false negative predictions as well as the decision boundary of your anomaly detection. How much variance is explained by the first two main components?

PROJECT 3:
BLOOD DONOR CLASSIFICATION

2. Explainable Model: For your decision support your model should be explainable. Train a model with a focus on explainability with an as simple as possible structure while still maintaining its predictive power.

- Train a **decision tree classifier** on the imputed data. Evaluate your model’s accuracy and visualize the tree structure to help the hospital personal understand the decision process. Each inference should not only put out **the class, but also the decision path taken**. Make the tree as **simple and understandable** as possible.

PROJECT 3:
BLOOD DONOR CLASSIFICATION

3. High Performance Model: This time the focus is on predictive power. Try and train a more accurate model. Is it worth the effort?

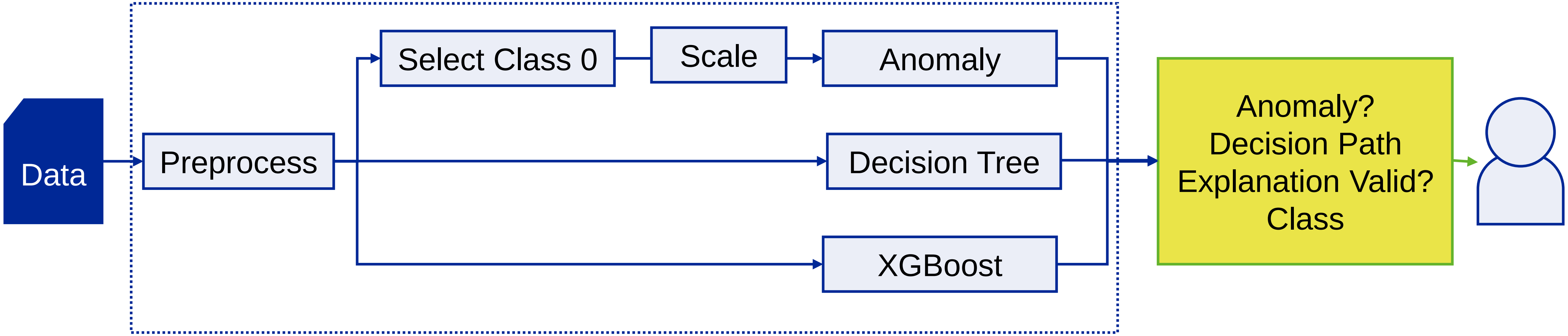
- **Train and optimize an XGBoost classifier** on the imputed data.
- Use **SHAP local explanation** techniques on 5 selected data points and discuss the results
- Use **SHAP global explanation** techniques to visualize and discuss the influence of different features.
- **Evaluate the XGBoost's accuracy** and compare it to the Decision Tree

PROJECT 3:
BLOOD DONOR CLASSIFICATION

4. Combined Model: Put all components into a single model artifact for deployment such that clinic personal has all important information at hand to make an informed decision.

> **Combine** the XGBoost, Decision Tree and Anomaly Detection in a **single model class** including all necessary methods (fit, predict…). The Decision Tree provides an explainable assistance for the hospital personal and the XGBoost (probably) a more accurate classification. The Anomaly Detection increases the robustness of the model for conditions that have not been explicitly trained or for human errors. **Generate a few test anomalies to check your detection**.

> **Evaluate, discuss and plot** the performance of your combined model.



PROJECT 3:
BLOOD DONOR CLASSIFICATION

- The submission consists of a resaonable amount filled in Jupyter Notebooks. There are **7 Points** for the complete and functional solution.
- There are **3 Points** for presentation, style, and creativity
- Due Date is **January 10th, 2025**

WHAT‘S TO BE LEARNED

- **Learn how to extract information from data**
  - Deal with different data types (structured and unstructured)
  - ML algorithms
  - Explore
  - Statistics
  - Linear algebra
  - Use analysis to find the sweet spots

WHAT‘S TO BE LEARNED

- **Learn how to interpret your findings**
  - What are the methods and metrics?
  - What correlations can be found?
  - What‘s significant and what‘s not?

WHAT‘S TO BE LEARNED

- **Communicate and present your results**
  - Make pretty graphs
  - Your data contains a story. Tell it!
  - Make mistaces, harness criticism, defend your theses
  - Convince your peers, boss, customer, client, referee …